



500.43451X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): KANO, et al.

Serial No.: 10/766,015

Filed: January 29, 2004

Title: DISK ARRAY SYSTEM AND METHOD FOR CONTROLLING DISK
ARRAY SYSTEM

LETTER CLAIMING RIGHT OF PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

February 25, 2004

Sir:

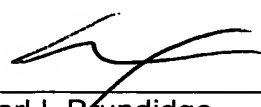
Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby
claim(s) the right of priority based on:

Japanese Patent Application No. 2003-400517
Filed: November 28, 2003

A certified copy of said Japanese Patent Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP



Carl I. Brundidge
Registration No.: 29,621

CIB/rr
Attachment

日 本 国 特 許 庁
JAPAN PATENT OFFICE

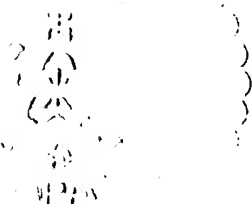
別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 3 年 1 1 月 2 8 日
Date of Application:

出 願 番 号 特 願 2 0 0 3 - 4 0 0 5 1 7
Application Number:
[ST. 10/C] : [J P 2 0 0 3 - 4 0 0 5 1 7]

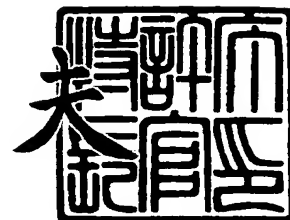
出 願 人 株式会社日立製作所
Applicant(s):



2 0 0 4 年 1 月 2 7 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康



出証番号 出証特 2 0 0 4 - 3 0 0 3 0 1 2

【書類名】 特許願
【整理番号】 340301442
【提出日】 平成15年11月28日
【あて先】 特許庁長官殿
【国際特許分類】 G06F 3/06
【発明者】
 【住所又は居所】 神奈川県小田原市中里 3 2 2 番 2 号 株式会社日立製作所 R A
 I D システム事業部内
 【氏名】 加納 東
【発明者】
 【住所又は居所】 神奈川県小田原市中里 3 2 2 番 2 号 株式会社日立製作所 R A
 I D システム事業部内
 【氏名】 小河 卓二
【発明者】
 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所
 システム開発研究所内
 【氏名】 八木沢 育哉
【特許出願人】
 【識別番号】 000005108
 【氏名又は名称】 株式会社日立製作所
【代理人】
 【識別番号】 110000176
 【氏名又は名称】 一色国際特許業務法人
 【代表者】 一色 健輔
【手数料の表示】
 【予納台帳番号】 211868
 【納付金額】 21,000円
【提出物件の目録】
 【物件名】 特許請求の範囲 1
 【物件名】 明細書 1
 【物件名】 図面 1
 【物件名】 要約書 1

【書類名】 特許請求の範囲**【請求項 1】**

第一のインタフェース規格によりデータの送受信を行う複数のハードディスクドライブが通信経路で接続され、当該複数のハードディスクドライブにより形成されている一つ又は複数の R A I D グループが収容されている第一の筐体と、

前記第一の筐体に収容されている前記ハードディスクドライブよりも信頼性の低い、第二のインタフェース規格によりデータの送受信を行う複数のハードディスクドライブが前記第一のインタフェース規格と前記第二のインタフェース規格とを変換する複数の変換装置を介して前記通信経路で接続され、当該複数のハードディスクドライブにより形成されている一つ又は複数の R A I D グループが収容されている第二の筐体と、

情報処理装置と通信可能に接続され、前記情報処理装置から前記第一の筐体または前記第二の筐体の前記ハードディスクドライブに対するデータの読み出し要求と書き込み要求とを受信するチャネル制御部と、前記通信経路を介して前記第一の筐体および前記第二の筐体の前記複数のハードディスクドライブと通信可能に接続され、前記チャネル制御部により受信される前記読み出し要求または前記書き込み要求をもとに、前記第一の筐体および前記第二の筐体の前記複数のハードディスクドライブとの間でデータおよび当該データを含む複数のデータに対する誤りを検出するためのデータであるパリティデータの入出力を行うディスク制御部と、前記複数のハードディスクドライブに書き込まれるデータを一時的に記憶するキャッシュメモリと、前記チャネル制御部と前記ディスク制御部との制御を司る C P U とを含んで構成されるコントローラと

を有し、

前記コントローラは、前記第二の筐体の前記複数のハードディスクドライブに記憶されているデータについて、当該データが記憶されている前記ハードディスクドライブが属している前記 R A I D グループの全ての前記ハードディスクドライブから、当該データを含む複数のデータと当該複数のデータに対するパリティデータとを読み出し、当該データを含む複数のデータが誤った内容で前記ハードディスクドライブに書き込まれていないか検査する

ことを特徴とするディスクアレイ装置。

【請求項 2】

請求項 1 に記載のディスクアレイ装置において、

前記コントローラは、

前記情報処理装置から前記第二の筐体の前記ハードディスクドライブに記憶されているデータの読み出し要求を受信すると、当該データについて前記検査を実施する

ことを特徴とするディスクアレイ装置。

【請求項 3】

請求項 1 に記載のディスクアレイ装置において、

前記コントローラは、

前記情報処理装置からの前記読み出し要求とは関係なく、前記第二の筐体の前記複数のハードディスクドライブに記憶されているデータについて前記検査を実施する

ことを特徴とするディスクアレイ装置。

【請求項 4】

請求項 1 に記載のディスクアレイ装置において、

前記コントローラは、

前記情報処理装置からの前記書き込み要求に従い前記第二の筐体の前記ハードディスクドライブに書き込んだデータについて、当該データが書き込まれている前記ハードディスクドライブにおける位置を更新管理テーブルに記憶し、

前記更新管理テーブルに記憶されている前記位置に記憶されている前記データについて、前記検査を実施し、当該位置に記憶されている前記データについて前記検査が完了したことを示す情報を前記更新管理テーブルに記憶し、

前記情報処理装置から前記第二の筐体の前記ハードディスクドライブに記憶されている

データの前記読み出し要求を受信すると、前記更新管理テーブルを参照し、当該データについて前記検査が実施されていない場合は当該データについて前記検査を実施することを特徴とするディスクアレイ装置。

【請求項 5】

請求項 1 に記載のディスクアレイ装置において、
前記コントローラは、

前記情報処理装置からの前記書き込み要求に従い前記第二の筐体の前記ハードディスクドライブにデータを書き込むと、当該ハードディスクドライブが備えるヘッドを当該データが記憶されている位置から移動した後で、当該データを前記ハードディスクドライブが備える磁気ディスクと前記キャッシュメモリとから読み出し、当該読み出した各々のデータを比較する

ことを特徴とするディスクアレイ装置。

【請求項 6】

請求項 1 に記載のディスクアレイ装置において、
前記コントローラは、

前記情報処理装置からの前記書き込み要求に従い前記第二の筐体の前記ハードディスクドライブにデータを書き込むと、当該ハードディスクドライブが備えるヘッドを当該データが記憶されている位置から移動した後で、当該データの一部を前記ハードディスクドライブが備える磁気ディスクと前記キャッシュメモリとから読み出し、当該読み出した各々のデータを比較する

ことを特徴とするディスクアレイ装置。

【請求項 7】

請求項 1 に記載のディスクアレイ装置において、
前記コントローラは、

前記情報処理装置からの前記書き込み要求に従い前記第二の筐体の前記ハードディスクドライブにデータを書き込むと、当該ハードディスクドライブが備えるヘッドを当該データが記憶されている位置から移動した後に、

当該データのサイズが所定の値未満である場合は、当該データを前記ハードディスクドライブが備える磁気ディスクと前記キャッシュメモリとから読み出し、当該読み出した各々のデータを比較し、

当該データのサイズが所定の値以上である場合は、当該データの一部を前記磁気ディスクと前記キャッシュメモリとから読み出し、当該読み出した各々のデータを比較する

ことを特徴とするディスクアレイ装置。

【請求項 8】

請求項 1 に記載のディスクアレイ装置において、
前記コントローラは、

前記情報処理装置からの書き込み要求が所定の数を超えた前記第二の筐体の前記ハードディスクドライブについて、当該ハードディスクドライブが備えるディスクキャッシュに記憶されているデータを当該ハードディスクドライブが備える磁気ディスクに書き込み、当該データを前記磁気ディスクと前記コントローラが備える前記キャッシュメモリとから読み出し、当該読み出した各々のデータを比較する

ことを特徴とするディスクアレイ装置。

【請求項 9】

請求項 1 に記載のディスクアレイ装置において、
前記コントローラは、

所定の時間ごとに前記第二の筐体の前記ハードディスクドライブについて、当該ハードディスクドライブが備えるディスクキャッシュに記憶されているデータを当該ハードディスクドライブが備える磁気ディスクに書き込み、当該データを前記磁気ディスクと前記コントローラが備える前記キャッシュメモリとから読み出し、当該読み出した各々のデータを比較する

ことを特徴とするディスクアレイ装置。

【請求項 10】

請求項 1 に記載のディスクアレイ装置において、
前記コントローラは、

前記第二の筐体の前記ハードディスクドライブにおいて、前記ハードディスクドライブが備えるディスクキャッシュの空領域が無くなると、当該ハードディスクドライブの前記ディスクキャッシュに記憶されているデータを当該ハードディスクドライブが備える磁気ディスクに書き込み、当該データを前記磁気ディスクと前記コントローラが備える前記キャッシュメモリとから読み出し、当該読み出した各々のデータを比較する

ことを特徴とするディスクアレイ装置。

【請求項 11】

請求項 1 に記載のディスクアレイ装置において、

前記ハードディスクドライブはデータの読み書きを行うヘッドを複数有し、

前記コントローラは、

前記情報処理装置からの前記書き込み要求に従い前記第二の筐体の前記ハードディスクドライブにデータの書き込みを行った前記ヘッドをヘッドチェック管理テーブルに記憶し

、
前記情報処理装置からの前記読み出し要求を受信すると、前記ヘッドチェック管理テーブルを参照し、当該データの読み出しに用いる前記ヘッドが前記ヘッドチェック管理テーブルに記憶されている場合は、当該ヘッドを用いて前記ハードディスクドライブが備える磁気ディスクに検査用のデータを書き込み、当該検査用のデータを前記磁気ディスクから読み出して前記検査用のデータと当該読み出したデータとを比較する

ことを特徴とするディスクアレイ装置。

【請求項 12】

請求項 1 に記載のディスクアレイ装置において、

前記コントローラは、

前記情報処理装置から前記第二の筐体の前記ハードディスクドライブに対するデータの書き込み要求を受信すると、前記データから複数セクタにより構成されるデータと当該複数セクタのデータの誤りを検出するためのデータであるパリティデータとでデータユニットを形成し、前記データユニットを前記ハードディスクドライブに書き込み、

前記情報処理装置から前記データの読み出し要求を受信すると、前記データユニットを読み出し、当該データが誤った内容で前記ハードディスクドライブに記憶されていないか検査する

ことを特徴とするディスクアレイ装置。

【請求項 13】

請求項 12 に記載のディスクアレイ装置において、

前記コントローラは、

前記データユニットを前記第二の筐体の前記 R A I D グループにおける複数のハードディスクドライブに分割して書き込む

ことを特徴とするディスクアレイ装置。

【請求項 14】

請求項 1 ～ 13 に記載のディスクアレイ装置において、

前記第一のインタフェース規格がファイバチャネルであり、前記第二のインタフェース規格がシリアル A T A であり、前記通信経路が F C - A L である

ことを特徴とするディスクアレイ装置。

【請求項 15】

請求項 1 ～ 13 に記載のディスクアレイ装置において、

前記第一のインタフェース規格がファイバチャネルであり、前記第二のインタフェース規格がパラレル A T A であり、前記通信経路が F C - A L である

ことを特徴とするディスクアレイ装置。

【請求項 16】

第一のインタフェース規格によりデータの送受信を行う複数のハードディスクドライブが通信経路で接続され、当該複数のハードディスクドライブにより形成されている一つ又は複数の R A I D グループが収容されている第一の筐体と、

前記第一のインタフェース規格の前記ハードディスクドライブよりも信頼性の低い第二のインタフェース規格によりデータの送受信を行う複数のハードディスクドライブが前記第一のインタフェース規格と前記第二のインタフェース規格とを変換する複数の変換装置を介して前記通信経路で接続され、当該複数のハードディスクドライブにより形成されている一つ又は複数の R A I D グループが収容されている第二の筐体と、

情報処理装置と通信可能に接続され、前記情報処理装置から前記第一の筐体または前記第二の筐体の前記ハードディスクドライブに対するデータの読み出し要求と書き込み要求とを受信するチャネル制御部と、前記通信経路を介して前記第一の筐体および前記第二の筐体の前記複数のハードディスクドライブと通信可能に接続され、前記チャネル制御部により受信される前記読み出し要求または前記書き込み要求をもとに、前記第一の筐体および前記第二の筐体の前記複数のハードディスクドライブとの間でデータおよび当該データを含む複数のデータに対する誤りを検出するためのデータであるパリティデータの入出力を行うディスク制御部と、前記複数のハードディスクドライブに書き込まれるデータを一時的に記憶するキャッシュメモリと、前記チャネル制御部と前記ディスク制御部との制御を司る C P U とを含んで構成されるコントローラと

を有するディスクアレイ装置の制御方法であって、

前記コントローラは、前記第二の筐体の前記複数のハードディスクドライブに記憶されているデータについて、当該データが記憶されている前記ハードディスクドライブが属している前記 R A I D グループの全ての前記ハードディスクドライブから、当該データを含む複数のデータと当該複数のデータに対するパリティデータとを読み出すステップと、

当該データを含む複数のデータが誤った内容で前記ハードディスクドライブに書き込まれていないか検査するステップと

を備えることを特徴とするディスクアレイ装置の制御方法。

【請求項 17】

請求項 16 に記載のディスクアレイ装置の制御方法において、

前記コントローラは、

前記情報処理装置からの前記書き込み要求に従い前記第二の筐体の前記ハードディスクドライブにデータを書き込むと、当該データを書き込んだ後に当該ハードディスクドライブが備えるヘッドを当該データが記憶されている位置から移動するステップと、

前記移動の後に当該データを前記ハードディスクドライブが備える磁気ディスクと前記キャッシュメモリとから読み出すステップと、

当該読み出した各々のデータを比較するステップと

を備えることを特徴とするディスクアレイ装置の制御方法。

【請求項 18】

請求項 16 に記載のディスクアレイ装置の制御方法において、

前記コントローラは、

前記情報処理装置から前記第二の筐体の前記ハードディスクドライブに対するデータの書き込み要求を受信すると、前記データから複数セクタにより構成されるデータと当該複数セクタのデータの誤りを検出するためのデータであるパリティデータとでデータユニットを形成するステップと、

前記データユニットを前記ハードディスクドライブに書き込むステップと、

前記情報処理装置から前記データの読み出し要求を受信すると、前記データユニットを読み出すステップと、

当該データが誤った内容で前記ハードディスクドライブに記憶されていないか検査するステップと

を備えることを特徴とするディスクアレイ装置の制御方法。

【請求項 1 9】

請求項 1 6 ～ 1 8 に記載のディスクアレイ装置の制御方法において、
前記第一のインタフェース規格がファイバチャネルであり、前記第二のインタフェース
規格がシリアル A T A であり、前記通信経路が F C - A L である
ことを特徴とするディスクアレイ装置の制御方法。

【請求項 2 0】

請求項 1 6 ～ 1 8 に記載のディスクアレイ装置の制御方法において、
前記第一のインタフェース規格がファイバチャネルであり、前記第二のインタフェース
規格がパラレル A T A であり、前記通信経路が F C - A L である
ことを特徴とするディスクアレイ装置の制御方法。

【書類名】 明細書**【発明の名称】** ディスクアレイ装置及びディスクアレイ装置の制御方法**【技術分野】****【0001】**

本発明は、ディスクアレイ装置及びディスクアレイ装置の制御方法に関する。

【背景技術】**【0002】**

近年、ディスクアレイ装置における記憶容量の増大に伴い、情報処理システムにおける重要性は益々高まってきている。そこで、情報処理装置等からのデータ入出力要求に対して、要求された位置に正しくデータの書き込みを行うこと、読み出したデータが不正である場合にはそれを検知することが重要である。

【0003】

特許文献1においては、磁気ディスク装置に2つのヘッドを持たせ、同一のデータを2つのヘッドから読み出して比較することにより、磁気ディスク装置における書き込み及び読み出しの信頼性を高める方法が開示されている。

【特許文献1】 特開平5-150909号公報

【発明の開示】**【発明が解決しようとする課題】****【0004】**

特許文献1の方法をディスクアレイ装置に適用する場合、各磁気ディスクにヘッドを2つ持たせる必要があるため、ハードディスクドライブの製造単価が高くなってしまう。そこで、ヘッドの追加等の物理的な構造の変更を行うことなく、ハードディスクドライブにおける信頼性を高める方法が求められている。

【0005】

また、ディスクアレイ装置においては、ファイバチャネルのハードディスクドライブに加えて、シリアルATAやパラレルATA等のハードディスクドライブ等も利用されはじめている。これは、シリアルATAやパラレルATA等のハードディスクドライブは、ファイバチャネルのハードディスクドライブと比較して信頼性は劣るが価格が低いためである。そこで、このようにファイバチャネルとシリアルATA等の規格のハードディスクドライブを組み合わせる構成するディスクアレイ装置において、ファイバチャネル以外のハードディスクドライブにおける信頼性を高める方法が求められている。

【0006】

本発明は上記課題を鑑みてなされたものであり、ディスクアレイ装置及びディスクアレイ装置の制御方法を提供することを目的とする。

【課題を解決するための手段】**【0007】**

上記目的を達成する本発明のうち主たる発明に係るディスクアレイ装置は、第一のインタフェース規格によりデータの送受信を行う複数のハードディスクドライブが通信経路で接続され、当該複数のハードディスクドライブにより形成されている一つ又は複数のRAIDグループが収容されている第一の筐体と、前記第一のインタフェース規格の前記ハードディスクドライブよりも信頼性の低い第二のインタフェース規格によりデータの送受信を行う複数のハードディスクドライブが前記第一のインタフェース規格と前記第二のインタフェース規格とを変換する複数の変換装置を介して前記通信経路で接続され、当該複数のハードディスクドライブにより形成されている一つ又は複数のRAIDグループが収容されている第二の筐体と、情報処理装置と通信可能に接続され、前記情報処理装置から前記第一の筐体または前記第二の筐体の前記ハードディスクドライブに対するデータの読み出し要求と書き込み要求とを受信するチャンネル制御部と、前記通信経路を介して前記第一の筐体および前記第二の筐体の前記複数のハードディスクドライブと通信可能に接続され、前記チャンネル制御部により受信される前記読み出し要求または前記書き込み要求をもとに、前記第一の筐体および前記第二の筐体の前記複数のハードディスクドライブとの間で

データおよび当該データを含む複数のデータに対する誤りを検出するためのデータであるパリティデータの入出力を行うディスク制御部と、前記複数のハードディスクドライブに書き込まれるデータを一時的に記憶するキャッシュメモリと、前記チャネル制御部と前記ディスク制御部との制御を司るCPUとを含んで構成されるコントローラとを有し、前記コントローラは、前記第二の筐体の前記複数のハードディスクドライブに記憶されているデータについて、当該データが記憶されている前記ハードディスクドライブが属している前記RAIDグループの全ての前記ハードディスクドライブから、当該データを含む複数のデータと当該複数のデータに対するパリティデータとを読み出し、当該データを含む複数のデータが誤った内容で前記ハードディスクドライブに書き込まれていないか検査する。

【0008】

また、前記コントローラは、前記情報処理装置からの前記書き込み要求に従い前記第二の筐体の前記ハードディスクドライブにデータを書き込むと、当該ハードディスクドライブが備えるヘッドを当該データが記憶されている位置から移動した後で、当該データを前記ハードディスクドライブが備える磁気ディスクと前記キャッシュメモリとから読み出し、当該読み出した各々のデータを比較する。

【0009】

また、前記コントローラは、前記情報処理装置から前記第二の筐体の前記ハードディスクドライブに対するデータの書き込み要求を受信すると、前記データから複数セクタにより構成されるデータと当該複数セクタのデータの誤りを検出するためのデータであるパリティデータとでデータユニットを形成し、前記データユニットを前記ハードディスクドライブに書き込み、前記情報処理装置から前記データの読み出し要求を受信すると、前記データユニットを読み出し、当該データが誤った内容で前記ハードディスクドライブに記憶されていないか検査する。

【0010】

ここで、第一のインタフェース規格とは例えばファイバチャネルのことであり、第二のインタフェース規格とは例えばシリアルATAのことであり、通信経路とは例えばFC-A Lのことであり。また、変換装置とは、例えばファイバチャネルプロトコルとシリアルATAプロトコルとを変換するコンバータのことであり。また、RAIDグループとは、ハードディスクドライブがRAID構成である場合において、複数のハードディスクドライブを1つのグループとして管理しているものである。RAIDグループ上には、情報処理装置からのアクセス単位である論理ボリュームが形成されており、各論理ボリュームにはLUNと呼ばれる識別子が付与されている。ディスク制御部は、情報処理装置から論理ボリュームに対するデータの書き込み要求を受信すると、当該データ及び当該データの誤りを検出するためのパリティデータ等をRAIDグループを形成するハードディスクドライブに対して書き込む。

【0011】

その他、本願が開示する課題、及びその解決方法は、発明を実施するための最良の形態の欄、及び図面により明らかにされる。

【発明の効果】

【0012】

ディスクアレイ装置及びディスクアレイ装置の制御方法を提供することができる。

【発明を実施するための最良の形態】

【0013】

==装置構成==

図1(a)は本発明の一実施例として説明するディスクアレイ装置10の正面図であり、図1(b)はディスクアレイ装置10の背面図である。図2(a)は、ディスクアレイ装置10に装着される基本筐体20を正面側から見た斜視図であり、図2(b)は基本筐体20を背面側から見た斜視図である。図3(a)は、ディスクアレイ装置10に装着される増設筐体30を正面側から見た斜視図であり、図3(b)は増設筐体30を背面側か

ら見た斜視図である。

【0014】

図1(a), (b)に示すように、ディスクアレイ装置10は、ラックフレーム11をベースとして構成される。ラックフレーム11の内側左右側面の上下方向には、複数段にわたって前後方向にマウントフレーム12が形成され、このマウントフレーム12に沿って基本筐体20および増設筐体30が引き出し式に装着される。図2(a), (b)に示すように、基本筐体20および増設筐体30には、ディスクアレイ装置10の各種機能を提供するボードやユニットが装着されている。

【0015】

図2(a)に示すように、基本筐体20の正面上段側には、ハードディスクドライブ51が装填された複数のディスクドライブユニット52が並べて装着されている。ハードディスクドライブ51は、例えば、FC-A L規格、SCSI1 (Small Computer System Interface 1) 規格、SCSI2規格、SCSI3規格、パラレルATA (AT Attachment) 規格、シリアルATA規格などの通信機能を提供する通信インタフェースを有するハードディスクドライブである。

【0016】

また、基本筐体20の正面下段側には、バッテリーユニット53、ハードディスクドライブ51の稼働状態などが表示される表示パネル54、フレキシブルディスクドライブ55が装着されている。バッテリーユニット53には二次電池が内蔵されている。バッテリーユニット53は、停電などによりAC/DC電源57からの電力供給が途絶えた場合に、ボードやユニットに電力を供給するバックアップ電源として機能する。表示パネル54には、ハードディスクドライブ51の稼働状態などを表示するLEDランプなどの表示デバイスが設けられている。フレキシブルディスクドライブ55は、メンテナンス用プログラムをロードする場合などに用いられる。

【0017】

図2(b)に示すように、基本筐体20の背面上段側の両側面側には、1枚ずつ電源コントローラボード56が装着されている。電源コントローラボード56は、複数のハードディスクドライブ51と通信可能に接続されている。電源コントローラボード56と複数のハードディスクドライブ51は、ループ状の通信経路、例えば、FC-A Lの方式(トポロジー)で通信を行う通信経路によって通信可能に接続されている。

【0018】

電源コントローラボード56は、AC/DC電源57の状態監視やハードディスクドライブ51の状態監視、ハードディスクドライブ51の電源供給の制御、冷却装置の冷却能力の制御、表示パネル54上の表示デバイスの制御、筐体各部の温度監視などを行う回路が実装されている。なお、冷却装置は、ディスクアレイ装置10内や筐体20, 30内を冷却する装置であり、例えば、インタークーラー、ヒートシンク、空冷式の冷却ファンなどである。電源コントローラボード56にはファイバチャネルケーブルのコネクタ67が設けられ、このコネクタにはファイバチャネルケーブル91が接続される。

【0019】

図2(b)に示すように、基本筐体20の背面上段側の前記2枚の電源コントローラボード56に挟まれた空間には、AC/DC電源57が2台並べて装着されている。AC/DC電源57は、ハードディスクドライブ51、ボード、ユニットなどに電源を供給する。AC/DC電源57は、電源コントローラボード56と接続されており、電源コントローラボード56からの信号により各ハードディスクドライブ51に電源を供給できるように設定されている。

【0020】

なお、本実施の形態においては、各筐体20, 30の電源供給に関するセキュリティを確保するために、基本筐体20および増設筐体30に電源コントローラボード56とAC/DC電源57とを各2台ずつ冗長に装着させることとしているが、電源コントローラボード56とAC/DC電源57とを各1台ずつ装着させることとしてもよい。

AC/DC電源57には、AC/DC電源57の出力をオン・オフするためのブレーカスイッチ64が設けられている。

【0021】

図2(b)に示すように、AC/DC電源57の下方には、2台の空冷式の冷却ファンユニット58が並べて装着されている。冷却ファンユニット58には、1台以上の冷却ファン66が実装されている。冷却ファン66は、筐体内に空気を流入・流出させることでハードディスクドライブ51やAC/DC電源57などから発生する熱を筐体外部に排出する。なお、基本筐体20や増設筐体30、およびこれらに装着されているボードやユニットには、筐体20、30内に空気を循環させる通気路や通気口が形成され、冷却ファン66により筐体20内の熱が外部に効率よく排出される仕組みになっている。冷却ファン66は、ハードディスクドライブ51ごとに設けることとしてもよいが、チップやユニットの数を削減できることから筐体ごとに大きな冷却ファン66を設けることが好ましい。

【0022】

冷却ファンユニット58は、コントローラボード59もしくは電源コントローラボード56と制御ライン48で接続されており、冷却ファンユニット58の冷却ファン66の回転数は、この制御ライン48を通じてコントローラボード59もしくは電源コントローラボード56により制御される。

【0023】

図2(b)に示すように、基本筐体20の背面下段側には、1枚のコントローラボード59が装着されている。コントローラボード59には、基本筐体20および増設筐体30に装着されているハードディスクドライブ51との間の通信インタフェースと、ハードディスクドライブ51の動作の制御（例えば、RAID方式による制御）やハードディスクドライブ51の状態監視を行う回路などが実装されている。

【0024】

なお、本実施の形態において、電源コントローラボード56がハードディスクドライブ51の電源供給の制御や冷却装置の冷却能力の制御を行うこととしているが、これらの制御をコントローラボード59が行うこととしてもよい。

【0025】

また、本実施例においては、コントローラボード59は、情報処理装置300との間の通信インタフェースの機能、例えば、SCSI規格やファイバチャネル規格の通信機能を提供する通信インタフェースボード61や、ハードディスクドライブ51への書き込みデータや読み出しデータが記憶されるキャッシュメモリ62などを実装する形態としているが、これらを別のボードが実装する形態としてもよい。

【0026】

コントローラボード59に装着される通信インタフェースボード61には、情報処理装置300と接続するための、ファイバチャネル、Ethernet（登録商標）などのプロトコルで構築されたSAN（Storage Area Network）、LAN（Local Area Network）、もしくは、SCSIなどの所定のインタフェース規格に準拠した外部コネクタ63が設けられ、ディスクアレイ装置10は、このコネクタ63に接続される通信ケーブル92を介して情報処理装置300と接続される。

【0027】

なお、基本筐体20のハードディスクドライブ51の制御に関するセキュリティを確保するために、2枚のコントローラボード59を冗長に装着させることとしてもよい。

【0028】

図3(a)に示すように、増設筐体30の正面側には、ハードディスクドライブ51が収容された複数のディスクドライブユニット52が並べて装着されている。図3(b)に示すように、増設筐体30の背面両側面側には、それぞれ一枚ずつ電源コントローラボード56が装着されている。また、2枚の電源コントローラボード56に挟まれた空間には、AC/DC電源57が2台並べて装着されている。また、AC/DC電源57の下方には、2台の冷却ファンユニット58が並べて装着されている。AC/DC電源57には、

AC/DC電源57の出力をオン・オフするためのブレーカスイッチ64が設けられている。

【0029】

本実施の形態においては、上述したように増設筐体30の電源供給に関するセキュリティを確保するために、増設筐体30に電源コントローラボード56とAC/DC電源57とを各2台ずつ冗長に装着させることとしているが、電源コントローラボード56とAC/DC電源57とを各1台ずつ装着させることとしてもよい。なお、ハードディスクドライブ51の電源供給の制御や冷却装置の冷却能力の制御などの電源コントローラボード56の機能をコントローラボード59に実装することとしてもよい。

【0030】

図4にディスクドライブユニット52に收容されているハードディスクドライブ51の構成の一例を示す。ハードディスクドライブ51は、その筐体70内に、磁気ディスク73、アクチュエータ71、スピンドルモータ72、データの読み書きを行うヘッド74、ヘッド74等の機構部分を制御する機構制御回路75、磁気ディスク73へのデータの読み書き信号を制御する信号処理回路76、通信インタフェース回路77、各種コマンドやデータが入出力されるインタフェースコネクタ79、電源コネクタ80等を備えて構成される。なお、通信インタフェース回路77には、データを一時的に格納するためのキャッシュメモリも含まれている。なお、後述するコントローラ500におけるキャッシュメモリ62と区別するため、ハードディスクドライブ51が備えるキャッシュメモリをディスクキャッシュと称する。

【0031】

ハードディスクドライブ51は、例えば、コンタクトスタートストップ(CSS: Contact Start Stop)方式の3.5インチサイズの磁気ディスクや、ロード/アンロード方式の2.5インチサイズの磁気ディスクなどを備える記憶装置である。3.5インチサイズの磁気ディスクは、例えば、SCSI1、SCSI2、SCSI3、FC-ALなどの通信インタフェースを有している。一方、2.5インチサイズの磁気ディスクは、例えば、パラレルATA、シリアルATAなどの通信インタフェースを有している。

【0032】

2.5インチサイズの磁気ディスクをディスクアレイ装置10の筐体20、30に收容する場合には、3.5インチの形状をした容器に収めるようにしてもよい。これにより、磁気ディスクの衝撃耐力性能を向上させることが可能となる。なお、2.5インチサイズの磁気ディスクと3.5インチサイズの磁気ディスクとは、通信インタフェースが異なるだけではなく、I/O性能、消費電力、寿命の点などで異なっている。2.5インチサイズの磁気ディスクは、3.5インチサイズの磁気ディスクに比べ、I/O性能が優れておらず、寿命が短い。しかし、3.5インチサイズの磁気ディスクに比べ、消費電力が少ないという点で優れている。

【0033】

==ディスクアレイ装置のハードウェア構成==

図5は、本発明の一実施例として説明するディスクアレイ装置10のハードウェア構成を示すブロック図である。

【0034】

図5に示すように、ディスクアレイ装置10には、SANを介して情報処理装置300が接続されている。情報処理装置300は、例えば、パーソナルコンピュータ、ワークステーション、メインフレームコンピュータなどである。

【0035】

ディスクアレイ装置10は、前述したように、基本筐体20と1つ又は複数の増設筐体30を備えている。本実施の形態においては、基本筐体20は、コントローラ500、ハードディスクドライブ51などを備えている。コントローラ500は、チャンネル制御部501、ディスク制御部502、CPU503、メモリ504、キャッシュメモリ62、及びデータコントローラ505などを備えており、前述のコントローラボード59に実装さ

れている。また、増設筐体30は、ハードディスクドライブ51などを備えている。基本筐体及び増設筐体のハードディスクドライブ51は、FC-A L 506により、ディスク制御部502と通信可能に接続されている。なお、ディスク制御部502とハードディスクドライブ51との接続形態の詳細については後述する。

【0036】

チャンネル制御部501は情報処理装置300との間で通信を行うインタフェースである。チャンネル制御部501は、ファイバチャネルプロトコルに従ってブロックアクセス要求を受け付ける機能を有する。

【0037】

ディスク制御部502は、CPU503からの指示により、ハードディスクドライブ51との間でデータのやりとりを行うインタフェースである。ディスク制御部502は、ハードディスクドライブ51を制御するコマンドなどを規定するプロトコルに従ってハードディスクドライブ51に対するデータ入出力要求を送信する機能を備える。

【0038】

CPU503は、ディスクアレイ装置10の全体の制御を司るもので、メモリ504に格納されたマイクロプログラムを実行することにより、チャンネル制御部501、ディスク制御部502、及びデータコントローラ506等の制御を行う。マイクロプログラムとは、図6に示すデータREAD処理601やデータWRITE処理602などである。

【0039】

キャッシュメモリ62は、チャンネル制御部501とディスク制御部502との間で授受されるデータを一時的に記憶するために用いられる。

【0040】

データコントローラ505は、CPU503の制御によりチャンネル制御部501とキャッシュメモリ62との間又はキャッシュメモリ62とディスク制御部502との間のデータ転送を行うものである。

【0041】

コントローラ500は、ハードディスクドライブ51をいわゆるRAID (Redundant Array of Inexpensive Disks) 方式に規定されるRAIDレベル (例えば、0, 1, 5) で制御する機能を備えている。RAID方式においては、複数のハードディスクドライブ51が1つのグループ (以後、RAIDグループと称する) として管理されている。RAIDグループ上には、情報処理装置300からのアクセス単位である論理ボリュームが形成されており、各論理ボリュームにはLUN (Logical Unit Number) と呼ばれる識別子が付与されている。RAIDの構成情報は、図6に示すようにメモリ504にRAID構成テーブル603として記憶されており、データREAD処理601やデータWRITE処理602の実行時にCPU503により参照される。

【0042】

なお、ディスクアレイ装置は、以上に説明した構成のもの以外にも、例えば、NFS (Network File System) などのプロトコルにより情報処理装置300からファイル名指定によるデータ入出力要求を受け付けるように構成されたNAS (Network Attached Storage) として機能するものなどであってもよい。

【0043】

==ハードディスクドライブの接続形態==

次に、コントローラ500とハードディスクドライブ51との接続形態について説明する。

図7は、基本筐体20にファイバチャネルのハードディスクドライブ51が収容されている場合における、ディスク制御部502と各ハードディスクドライブ51との接続形態を示している。

【0044】

ディスク制御部502は、FC-A L 506で複数のハードディスクドライブ51と接続されている。FC-A L 506は、複数のPBC (Port Bypass Circuit) 602を備

えている。ファイバチャネルのハードディスクドライブ51は、このPBC701を介してFC-AL506に接続されている。PBC701は、チップ化された電子スイッチであり、ディスク制御部502やハードディスクドライブ51などをバイパスし電氣的にFC-AL506から除外する機能も有している。具体的には、PBC701は、障害が発生したハードディスクドライブ51をFC-AL506から切り離して、他のハードディスクドライブ51とディスク制御部502との間の通信を可能にする。

【0045】

また、PBC701は、FC-AL506の動作を維持したままでハードディスクドライブ51の抜き差しを可能にする。例えば、ハードディスクドライブ51が新たに装着された場合にはそのハードディスクドライブ51をFC-AL506に取り込み、ディスク制御部502との間の通信を可能にする。なお、PBC701の回路基板は、ディスクアレイ装置10のラックフレーム11に設けられているか、もしくは、その一部または全部がコントローラボード59や電源コントローラボード56に実装されていることとしてもよい。

【0046】

図8は、基本筐体20にシリアルATAのハードディスクドライブ51が収容されている場合における、ディスク制御部502と各ハードディスクドライブ51との接続形態を示している。

各ハードディスクドライブ51は、コンバータ801を介してFC-AL506のPBC701に接続されている。コンバータ801はファイバチャネルプロトコルとシリアルATAプロトコルとを変換する回路である。コンバータ801は、プロトコル変換機能が組み込まれた1つのチップであり、各ディスクドライブユニット52内に設けられている。

【0047】

図9は、基本筐体20にシリアルATAのハードディスクドライブ51が収容されている場合における、もう一つの接続形態を示している。

コンバータ901は、図7におけるコンバータ801と同様にファイバチャネルプロトコルとシリアルATAプロトコルとを変換する回路である。コンバータ901はFC-AL506のPBC701に接続されており、1つのコンバータ901には、複数のハードディスクドライブ51がスイッチ902を介して接続されている。スイッチ902は、ハードディスクドライブ51が複数のコンバータ901に接続されている場合において、どのハードディスクドライブ51と通信を行うかを選択する回路である。スイッチ902は、各ディスクドライブユニット52内に設けられている。コンバータ901は、プロトコル変換機能が組み込まれた1つのチップであるか、複数の回路により構成されている。コンバータ901は、例えば「米国特許出願公開第2003/013557号明細書」にて開示されているSATAマスタデバイスの構成により実現することが可能である。コンバータ901は、コントローラボード59や電源コントローラボード56等の実装されている。

【0048】

==信頼性を高めるための制御==

以上に説明したディスクアレイ装置10において、ハードディスクドライブ51からの読み出し又はハードディスクドライブ51への書き込みの信頼性を高める方法について説明する。

【0049】

==RAID構成でのパリティチェック==

まず、RAID構成においてハードディスクドライブ51に記憶されているデータが不正な状態となっていないか検査する方法について説明する。ここで、不正な状態とは、データがディスク制御部502から指定された場所に指定された内容で書き込まれていない状態のことである。

【0050】

図10は、RAID5においてハードディスクドライブ51にデータが記憶されている様子を表している。RAID5においては、複数のハードディスクドライブ51によりRAIDグループ1001が形成されている。図10の例では、ハードディスクドライブ51には、データA～DとデータA～Dに対する誤りを検出するためのパリティデータP(A～D)が記憶されている。同様に、データE～HとデータE～Hに対するパリティデータP(E～H)が記憶されている。このような、データとパリティデータとの組合せのことを、ストライプグループ1002と呼ぶこととする。このようなストライプグループ1002が形成されているRAID構成において、コントローラ500はストライプグループ1002の全てのデータ及びパリティデータを読み出すことにより、データが不正な状態となっていないか検査することができる。まず、CPU503からの指示により、ディスク制御部502がデータA～DとパリティデータP(A～D)とを読み出す。次に、CPU503は、データA～DとパリティデータP(A～D)とでパリティチェックを行うことにより、データA～Dのいずれかが不正な状態となっていないか検査することができる。

【0051】

コントローラ500は、情報処理装置300からデータの読み出し要求を受信した際に、当該データを含むストライプグループの全てのデータとパリティデータとを読み出すようにすることもできる。これにより、コントローラ500がハードディスクドライブ51から不正なデータを読み出して情報処理装置300に送信することを防止することが可能となる。なお、不正なデータの検査はデータの読み出し要求を受信した際にかかわらず、任意のタイミングで行うこととしてもよい。これにより、データの読み出し性能に影響を与えずに、不正なデータの検出を行うことが可能となる。

【0052】

また、図11に示す更新管理テーブル1101を用いて、ハードディスクドライブ51に書き込まれたデータが不正な状態となっていないか検査することができる。更新管理テーブル1101はドライブ番号とセクタ番号とで構成され、メモリ504に記憶されている。本実施の形態においてはセクタ番号はLBA(Logical Block Address)で定義されており、LBA#1～128のように128LBA単位で管理されている。なお、セクタ番号をまとめる単位は128に限らず任意の単位でよい。CPU503は、ディスク制御部502を介してデータをハードディスクドライブ51に書き込むと、更新管理テーブル1101の当該ハードディスクドライブ51の書き込みを行ったセクタの値を「1」に変更する。CPU503は、更新管理テーブル1101において「1」が記憶されているハードディスクドライブ51の対象セクタのストライプグループの全てのデータとパリティデータとをディスク制御部502を介して読み出し、パリティチェックを行う。CPU503は、読み出したデータが不正でない場合は、更新管理テーブル1101において当該セクタの値を「0」に変更する。CPU503は、チャンネル制御部501を介して情報処理装置300からデータの読み出し要求を受信すると、更新管理テーブル1101を参照して当該データが記憶されているセクタが検査済みであるかどうかを確認する。当該データが記憶されているセクタが検査済みでない場合は、CPU503は前述の手順に従い当該データの属するストライプグループのデータを検査する。このように、ハードディスクドライブ51に書き込まれたデータについて、当該データに対する読み出し要求を受信する前に当該データの検査を実施しておくことにより、データの読み出し性能の低下を抑えることが可能となる。また、検査の未済を更新管理テーブル1101に記憶し、検査が行われていないデータを読み出す際にはパリティチェックを行うため、不正データの読み出しを防止することが可能となる。

【0053】

==WRITEデータに対する検査==

次に、ハードディスクドライブ51にデータを書き込んだ際に、当該データが正しく書き込まれているか検査する方法について説明する。

【0054】

図12は、コントローラ500がハードディスクドライブ51にデータを書き込む際のCPU503での制御を表すフローチャートである。CPU503は、チャンネル制御部501を介して情報処理装置300からデータの書き込み要求を受信すると、当該データのハードディスクドライブ51への書き込み指示をディスク制御部502に送信する(S1201)。そして、CPU503は当該データが書き込まれた磁気ディスクのヘッドの位置を移動させるシーク処理の実行指示をディスク制御部に送信する(S1202)。次に、CPU503は、キャッシュメモリ62から当該データを読み出し(S1203)、磁気ディスクから当該データを読み出す(S1204)。CPU503は、キャッシュメモリ62のデータと磁気ディスクのデータとが一致しているか比較する(S1205)。2つのデータが一致していない場合、CPU503は、書き込みが正常に行われていないことを情報処理装置300に通知する(S1206)。

このように、磁気ディスクに記憶されているデータとキャッシュメモリ62に記憶されているデータとを比較することにより、磁気ディスクに正しくデータが書き込まれているか確認することが可能である。また、書き込まれているデータが不正な状態となっている場合においても、データがキャッシュメモリ62に残っているため、データを失うことがない。なお、比較するデータを磁気ディスクとキャッシュメモリ62とから読み出す前に、シーク処理等により当該ハードディスクドライブが備えるヘッドを移動することにより、書き込み時にヘッドの位置が不正であった場合に、同じ位置から再度読み出すことを防止することが可能となる。

【0055】

図12の処理においては、書き込まれたデータの全部をキャッシュメモリ62と磁気ディスクとから読み出し比較することによりデータの検査を実施したが、データの全部ではなく、先頭と末尾の1セグメント等、そのデータの一部を読み出して比較することとしてもよい。例えば、シリアルATAのハードディスクドライブは、データのバックアップ等の用途に用いられるため、サイズの大きいデータ(シーケンシャルデータ)が書き込まれることが多い。このような場合、書き込みを行ったデータの全部について、磁気ディスクに記憶されているデータとキャッシュメモリ62に記憶されているデータとを比較することは、書き込み処理の性能を著しく低下させる要因となる。また、シーケンシャルデータの書き込み時に書き込み位置の誤り等が発生した場合は、そのデータの全部が不正となっている可能性が高い。そのため、データの一部を検査することでデータが不正となっているか判断可能である場合が多い。つまり、書き込みを行ったデータの一部、例えば先頭と末尾の1セグメント等について比較することにより、書き込み処理の性能低下を抑えうえて、不正データのチェックを行うことが可能である。

【0056】

また、ハードディスクドライブ51に書き込まれたデータのサイズに応じて、データの検査方法を変更することとしてもよい。図13は、書き込まれたデータがシーケンシャルデータであるかどうかに応じて、検査方法を変更する処理を示すフローチャートである。CPU503は、ハードディスクドライブ51にデータを書き込む指示をディスク制御部502に送信する(S1301)。そして、CPU503は当該データが書き込まれた磁気ディスクのヘッドの位置を移動させるシーク処理の実行指示をディスク制御部に送信する(S1302)。CPU503は、当該データがシーケンシャルデータであるかどうか判断する(S1303)。なお、シーケンシャルデータであるかどうかの判断は、書き込まれたデータのサイズが既定のサイズ以上であるかどうかにより行う。

CPU503は、当該データがシーケンシャルデータである場合は、キャッシュメモリ62と磁気ディスクとから当該データの先頭と末尾の1セグメントを読み出す。また、CPU503は、当該データがシーケンシャルデータでない場合は、キャッシュメモリ62と磁気ディスクとから当該データの全部を読み出す(S1306, S1307)。その後、CPU503は読み出した2つのデータが一致しているか比較し(S1308)、一致していない場合は書き込みが正常に行われていないことを情報処理装置300に通知する(S1309)。

このように、書き込んだデータがシーケンシャルデータである場合は、当該データの一部について、磁気ディスクとキャッシュメモリ 62 とに記憶されているデータを比較することにより、書き込み処理の性能低下を抑えた上で、データの不正を検出することが可能である。また、書き込んだデータがシーケンシャルデータでない場合は、書き込んだデータの全部について、磁気ディスクとキャッシュメモリ 62 とに記憶されているデータを比較することにより、シーケンシャルデータの場合ほど書き込み処理の性能を著しく低下させることなく、データの不正を完全に検出することが可能である。

【0057】

ハードディスクドライブ 51 は、データの書き込み性能を向上させるため、コントローラ 500 からデータの書き込み要求を受信すると、当該データをディスクキャッシュのみに書き込み、コントローラ 500 に書き込み完了を通知する機能を備えている場合がある。この場合、図 12 および図 13 に説明した方法では、書き込まれたデータの検査を行うことができない。図 14 は、ハードディスクドライブ 51 がこのような機能を備えている場合において、書き込まれたデータの検査を行う処理のフローチャートを示す図である。CPU 503 は、ハードディスクドライブ 51 への書き込み回数が所定の回数を超過していないか監視している (S1401)。所定の回数を超過すると、CPU 503 はディスクキャッシュに記憶されているデータを磁気ディスクに書き込むよう、ディスク制御部 4502 を介してハードディスクドライブに通知する (S1402)。そして、CPU 503 は、当該データをキャッシュメモリ 62 と磁気ディスクとから読み出す (S1403, S1404)。CPU 503 は、キャッシュメモリ 62 のデータと磁気ディスクのデータとが一致しているか確認し (S1405)、一致していない場合は書き込みが正常に行われていないことを情報処理措置 300 に通知する (S1406)。これにより、前述した書き込み処理の性能を高める機能を用いた上で、データの不正を検出することが可能となる。なお、図 14 の処理では、書き込み回数が所定の回数を超過した場合に磁気ディスクへのデータの書き込みと書き込まれたデータの検査を行うこととしたが、所定の時間が経過した場合や、ディスクキャッシュの空領域が無くなった場合などを契機としてもよい。

【0058】

また、シリアル ATA のハードディスクドライブ 51 においては、ヘッドの障害により、データの書き込みが正しく行われていないことが多い。そこで、ハードディスクドライブ 51 からデータを読み出す際に、ヘッドの障害を検出する方法を説明する。

【0059】


図 15 はヘッドチェック管理テーブル 1501 を示す図である。ヘッドチェック管理テーブル 1501 はドライブ番号とヘッド番号とセクタ番号とで構成され、メモリ 504 に記憶されている。セクタ番号は更新管理テーブル 1101 と同様に LBA により定義されている。CPU 503 は、ディスク制御部 502 を介してデータをハードディスクドライブ 51 に書き込むと、ヘッドチェック管理テーブル 1501 の当該データの書き込みを行ったヘッドの当該セクタの「更新有無」の値を「1」に変更する。

【0060】

図 16 は、CPU 503 が実行するヘッドチェック処理のフローチャートを示す図である。CPU 503 は、検査ヘッド番号に初期値として 1 を設定する (S1601)。CPU 503 は、一定時間経過するのを待ち (S1602)、検査ヘッド番号で指定されるヘッドを用いて磁気ディスクの管理領域に検査用のデータを書き込む (S1603)。なお管理領域は磁気ディスク上の予め定められている記憶領域である。次に CPU 503 は、管理領域に書き込まれているデータを読み出し (S1604)、読み出したデータと検査用のデータとが一致しているか確認する (S1605)。

【0061】

データが一致している場合、CPU 503 は当該ヘッドに異常が無いと判断し、ヘッドチェック管理テーブル 1501 の当該ヘッドの「更新有無」を「0」に変更する (S1606)。CPU 503 は、検査ヘッド番号に 1 を加算する (S1607)。検査ヘッド番号がヘッド番号の最大値より大きいか確認し (S1608)、大きい場合は検査ヘッド番



号を1に設定する。CPU503は、設定されたヘッド番号について、ヘッドチェック処理を繰り返し実行する。

【0062】

管理領域から読み出したデータと検査用のデータとが一致していない場合、CPU503は当該ハードディスクドライブ51に異常が発生していることを情報処理装置300に通知し、処理を終了する。

【0063】

図17は、CPU503が情報処理装置300からデータの読み出し要求を受信した際の処理のフローチャートを示す図である。CPU503は、チャンネル制御部501を介して情報処理装置300からデータの読み出し要求を受信する(S1701)。CPU503は、ヘッドチェック管理テーブル1501から、当該データが記憶されているハードディスクドライブ51の対象セクタの「更新有無」を確認する(S1702, S1703)。「更新有無」が「1」である場合は、当該ハードディスクドライブ51の当該LBAについてデータの書き込みが行われているが、前述したヘッドチェック処理が行われていない状態を示している。「更新有無」が「0」である場合は、CPU503は当該データをハードディスクドライブ51から読み出す(S1708)。

【0064】

「更新有無」が「1」である場合、CPU503は前述したヘッドチェック処理と同様に、当該ヘッドを用いて磁気ディスクの管理領域に検査用のデータを書き込む(S1704)。なお管理領域は磁気ディスク上の予め定められている記憶領域である。次にCPU503は、管理領域に書き込まれているデータを読み出し(S1705)、読み出したデータと検査用のデータとが一致しているか確認する(S1706)。

データが一致している場合、CPU503は当該ヘッドに異常が無いと判断し、ヘッドチェック管理テーブル1501の当該ヘッドの「更新有無」を「0」に変更する(S1707)。そして、CPU503は当該読み出し要求に従いハードディスクドライブ51からデータを読み出す(S1708)。

管理領域から読み出したデータと検査用のデータとが一致していない場合、CPU503は当該ハードディスクドライブ51に異常が発生していることを情報処理装置300に通知し(S1709)、当該ハードディスクドライブ51からデータを読み出さずに処理を終了する。

【0065】

これにより、ハードディスクドライブ51に書き込まれているデータを読み出す際に、当該データの書き込みを行ったヘッドが正常であるかどうかを確認することができる。ヘッドが異常である場合、データが正しく書き込まれていない可能性や、データの読み出しを正しく行うことができない可能性がある。データの読み出し時にヘッドの異常を検知することにより、不正なデータを読み出すことを防止することが可能となる。

【0066】

==パリティ付与による検査==

前述したRAID構成でのストライプグループの全てのデータを読み出してパリティチェックする方法では、ストライプグループの中のどのデータが不正な状態となっているか判断することができなかった。そのため、不正なデータの読み出しを防止することは可能であるが、不正なデータを復旧させることができず、データを損失してしまう可能性がある。そこで、ストライプグループにおけるパリティデータとは別に、各データにパリティデータを付与する方法について説明する。

【0067】

CPU503は、データをハードディスクドライブ51に書き込む際の最小単位を必ず複数セクタとし、これら複数セクタに対する誤りを検出するためのパリティデータを生成する。本実施の形態において、この複数セクタのデータとパリティデータとの組合せをデータユニットと称することとする。CPU503は、チャンネル制御部501を介して情報処理装置300からデータの書き込み要求を受信すると、当該データからデータユニット

を形成する。CPU 503は当該データユニットをディスク制御部502を介してハードディスクドライブ51に書き込む。

図18は、ハードディスクドライブに1つのデータ1801が書き込まれている様子を示す図である。データ1801は複数のセクタS#1～S#4で構成され、これら複数セクタのデータ1801に対するパリティデータ1802とでデータユニット1803が形成されている。CPU 503はチャンネル制御部501を介して情報処理装置300からデータの読み出し要求を受信すると、ディスク制御部502を介して当該データのデータユニット1803を読み出し、当該データのパリティチェックを行うことで当該データが不正な状態となっていないか検査する。このように、読み出し要求の対象となっているデータのみを読み出して、当該データが不正な状態となっていないか判断することが可能となる。また、ハードディスクドライブ51がRAID5のように冗長性のあるRAID構成である場合には、ストライプグループにおける他のデータ及びパリティデータとを用いて、当該データを復元することが可能であるため、データを損失することがない。

【0068】

1つのハードディスクドライブ51においてヘッ드의障害等が発生している場合は、不正な状態となっているセクタが複数発生する可能性が高い。データユニット1803が1つのハードディスクドライブ51に書き込まれている場合にデータユニット1803のうちの複数のセクタが不正な状態となると、パリティチェックにより不正を検出できない場合がある。

【0069】

そこで、図19に示すように、CPU 503は、前述したデータユニット1803をディスク制御部502を介してRAIDグループ内の複数のハードディスクドライブ51に分散して書き込むこととしてもよい。図20は、データユニット管理テーブル2001を示している図である。データユニット管理テーブル2001は、複数セクタで構成されるデータユニット1803がどのハードディスクドライブ51のどのLBAに対応しているかを示している。図20の例では、000～129までの130セクタで1つのデータユニット1803が形成され、このデータユニットはドライブ番号#0とドライブ番号#1のハードディスクドライブ51の000～064までのLBAにより構成されていることが示されている。CPU 503は、情報処理装置300からデータの書き込み要求を受信すると、データユニット管理テーブル2001を参照し、データユニット1803ごとに複数のハードディスクドライブ51に分散してデータを書き込む。

【0070】

これにより、1つのハードディスクドライブに障害が発生している場合においても、データの不正を検出することができる可能性が高くなる。

【0071】

==ファイバチャネルとシリアルATAとが混在する環境==

次に、ファイバチャネルのハードディスクドライブ51とシリアルATAのハードディスクドライブ51とが混在しているディスクアレイ装置10における説明を行う。

【0072】

図21は、第一の筐体2101にファイバチャネルのハードディスクドライブ51、第二の筐体2102にシリアルATAのハードディスクドライブ51が収容されているディスクアレイ装置を示すブロック図である。なお、第一の筐体2101及び第二の筐体2102とは、基本筐体20または増設筐体30のことである。各ハードディスクドライブ51のディスク制御部502との接続形態については、前述した通りである。また、図19においては、1つのコンバータ901に複数のシリアルATAのハードディスクドライブが接続される形態を示しているが、前述したように、コンバータ801が各ディスクドライブユニットに設けられて接続されているものとしてもよい。

【0073】

このようなディスクアレイ装置10においては、ファイバチャネルのハードディスクドライブ51と比較して信頼性の低いシリアルATAのハードディスクドライブ51の信頼

性を高めることが求められている。そこで、コントローラ 500 は、シリアル ATA のハードディスクドライブ 51 のみに対して、前述した信頼性を高める方法を適用する。これにより、基幹業務等の高いアクセス性能が要求される処理に用いられるファイバチャネルのハードディスクドライブ 51 に対するデータの読み書き性能を落とさずに、シリアル ATA のハードディスクドライブ 51 に対するデータの読み書きの信頼性を高めることができる。また、シリアル ATA のハードディスクドライブ 51 の各磁気ディスクにヘッドを 2 つずつ設ける等の物理的な構造の変更が必要でないため、シリアル ATA のハードディスクドライブ 51 の製造コストを抑えることが可能である。

【0074】

なお、本実施の形態においては、ファイバチャネルのハードディスクドライブ 51 とシリアル ATA のハードディスクドライブ 51 とが混在しているとしたが、信頼性の異なるインタフェース規格のハードディスクドライブ 51 であれば他のものでもよい。例えば、シリアル ATA のハードディスクドライブ 51 の代わりにパラレル ATA のハードディスクドライブ 51 であるとしてもよい。

【0075】

以上、本実施の形態について説明したが、上記実施例は本発明の理解を容易にするためのものであり、本発明を限定して解釈するためのものではない。本発明は、その趣旨を逸脱することなく、変更、改良され得ると共に、本発明にはその等価物も含まれる。

【図面の簡単な説明】

【0076】

【図 1】 本実施の形態における、ディスクアレイ装置の外観を示す図である。

【図 2】 本実施の形態における、ディスクアレイ装置の基本筐体の構成を示す図である。

【図 3】 本実施の形態における、ディスクアレイ装置の増設筐体の構成を示す図である。

【図 4】 本実施の形態における、ハードディスクドライブの構成を示す図である。

【図 5】 本実施の形態における、ディスクアレイ装置の構成を示す図である。

【図 6】 本実施の形態における、コントローラの CPU が実行するマイクロプログラムがメモリに記憶されている状態を示す図である。

【図 7】 本実施の形態における、ファイバチャネルのハードディスクドライブをコントローラのディスク制御部と接続する形態を示す図である。

【図 8】 本実施の形態における、シリアル ATA のハードディスクドライブをコントローラのディスク制御部と接続する第一の形態を示す図である。

【図 9】 本実施の形態における、シリアル ATA のハードディスクドライブをコントローラのディスク制御部と接続する第二の形態を示す図である。

【図 10】 本実施の形態における、RAID グループを構成するハードディスクドライブにデータが書き込まれている例を示す図である。

【図 11】 本実施の形態における、更新管理テーブルを示す図である。

【図 12】 本実施の形態における、データ書き込み時にキャッシュメモリと磁気ディスクとに記憶されているデータを比較するフローチャートを示す図である。

【図 13】 本実施の形態における、データ書き込み時にデータサイズを考慮してキャッシュメモリと磁気ディスクとに記憶されているデータを比較するフローチャートを示す図である。

【図 14】 本実施の形態における、ディスクキャッシュに記憶されているデータを磁気ディスクに書き込む際にキャッシュメモリと磁気ディスクとに記憶されているデータを比較するフローチャートを示す図である。

【図 15】 本実施の形態における、ヘッドチェック管理テーブルを示す図である。

【図 16】 本実施の形態における、定期的に実施するヘッドチェックのフローチャートを示す図である。

【図 17】 本実施の形態における、データの読み出し時にヘッドチェックを実施する

フローチャートを示す図である。

【図18】本実施の形態における、データユニットが1つのハードディスクドライブに書き込まれている例を示す図である。

【図19】本実施の形態における、データユニットが複数のハードディスクドライブに分散して書き込まれている例を示す図である。

【図20】本実施の形態における、データユニット管理テーブルを示す図である。

【図21】本実施の形態における、第一の筐体にファイバチャネルのハードディスクドライブが収容され、第二の筐体にシリアルATAのハードディスクドライブが収容されているディスクアレイ装置の構成を示す図である。

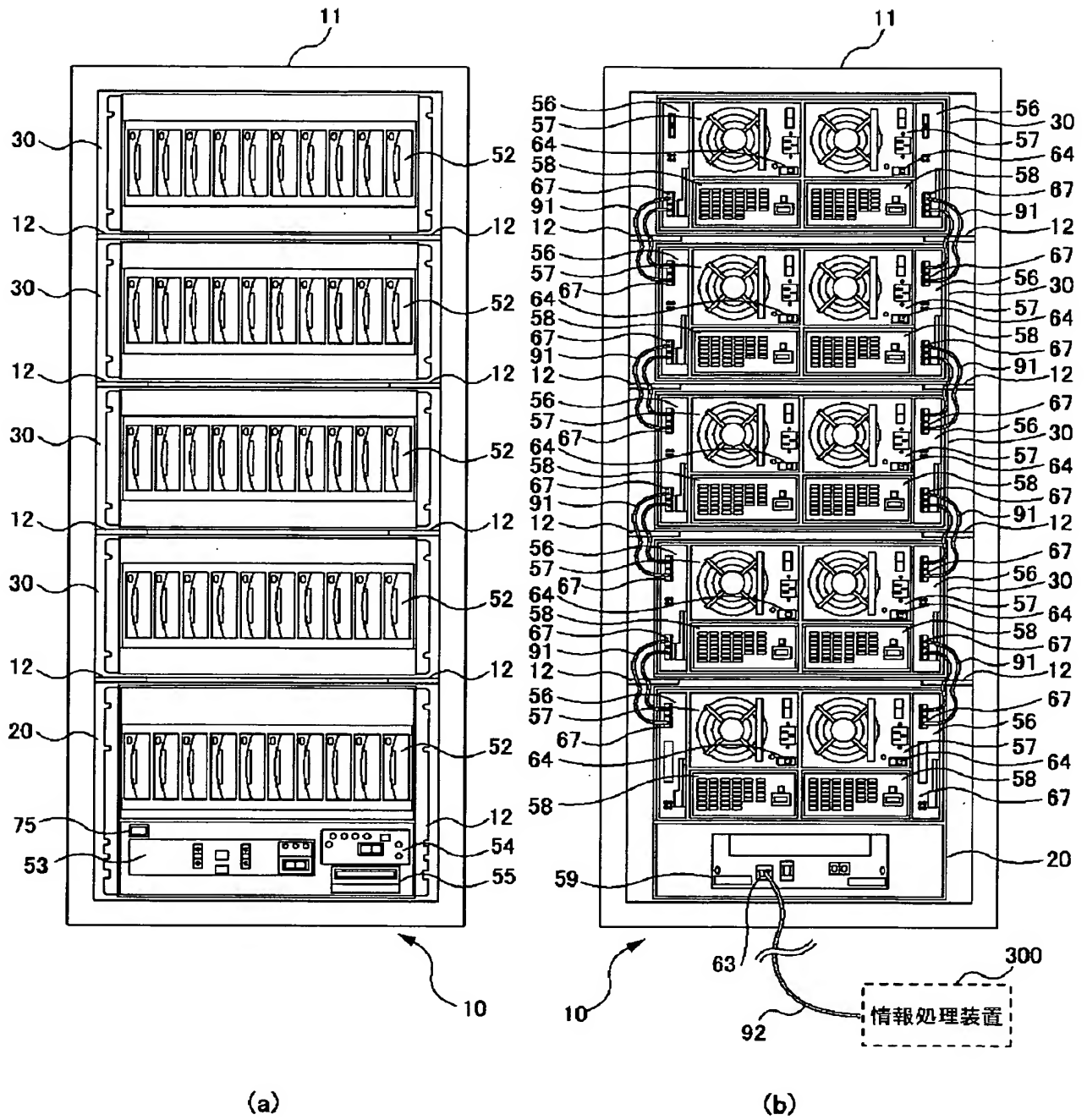
【符号の説明】

【0077】

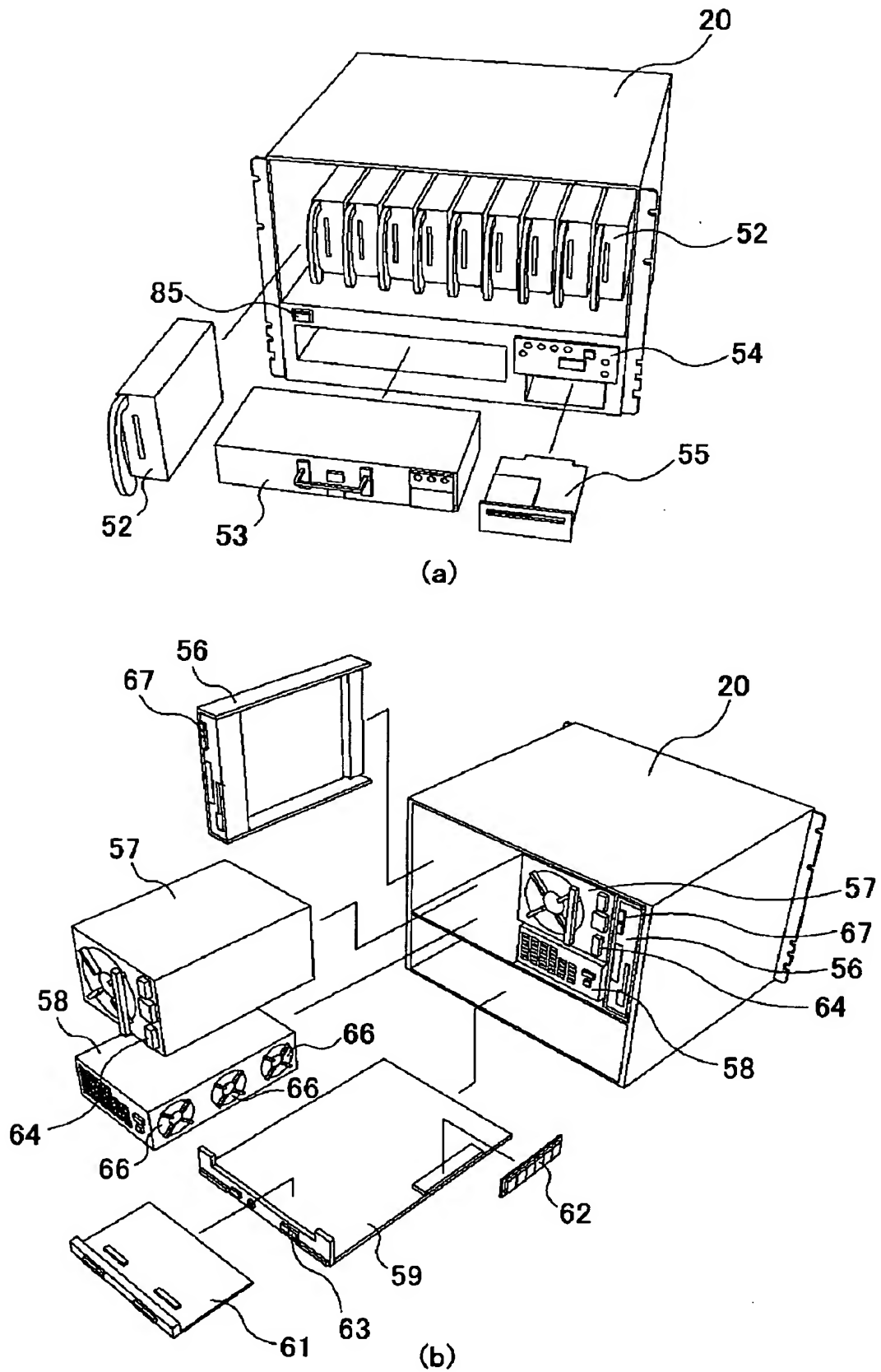
10	ディスクアレイ装置	11	ラックフレーム
12	マウントフレーム	20	基本筐体
30	増設筐体	48	制御ライン
49	電源供給ライン	51	ハードディスクドライブ
52	ディスクドライブユニット	56	電源コントローラボード
57	AC/DC電源	58	冷却ファンユニット
59	コントローラボード	61	通信インターフェースボード
62	キャッシュメモリ	64	ブレーカスイッチ
66	冷却ファン	67	コネクタ
70	ディスクドライブの筐体	73	磁気ディスク
85	メインスイッチ	81	電源コントローラ
91	ファイバチャネルケーブル	300	情報処理装置
500	コントローラ	501	チャネル制御部
502	ディスク制御部	503	CPU
504	メモリ	505	データコントローラ
506	FC-AL	701	PBC
801	コンバータ	901	コンバータ
902	スイッチ	1001	RAIDグループ
1002	ストライプグループ	1101	更新管理テーブル
1501	ヘッドチェック管理テーブル		
1801	複数セクタのデータ	1802	パリティデータ
1803	データユニット	2001	データユニット管理テーブル

【書類名】 図面

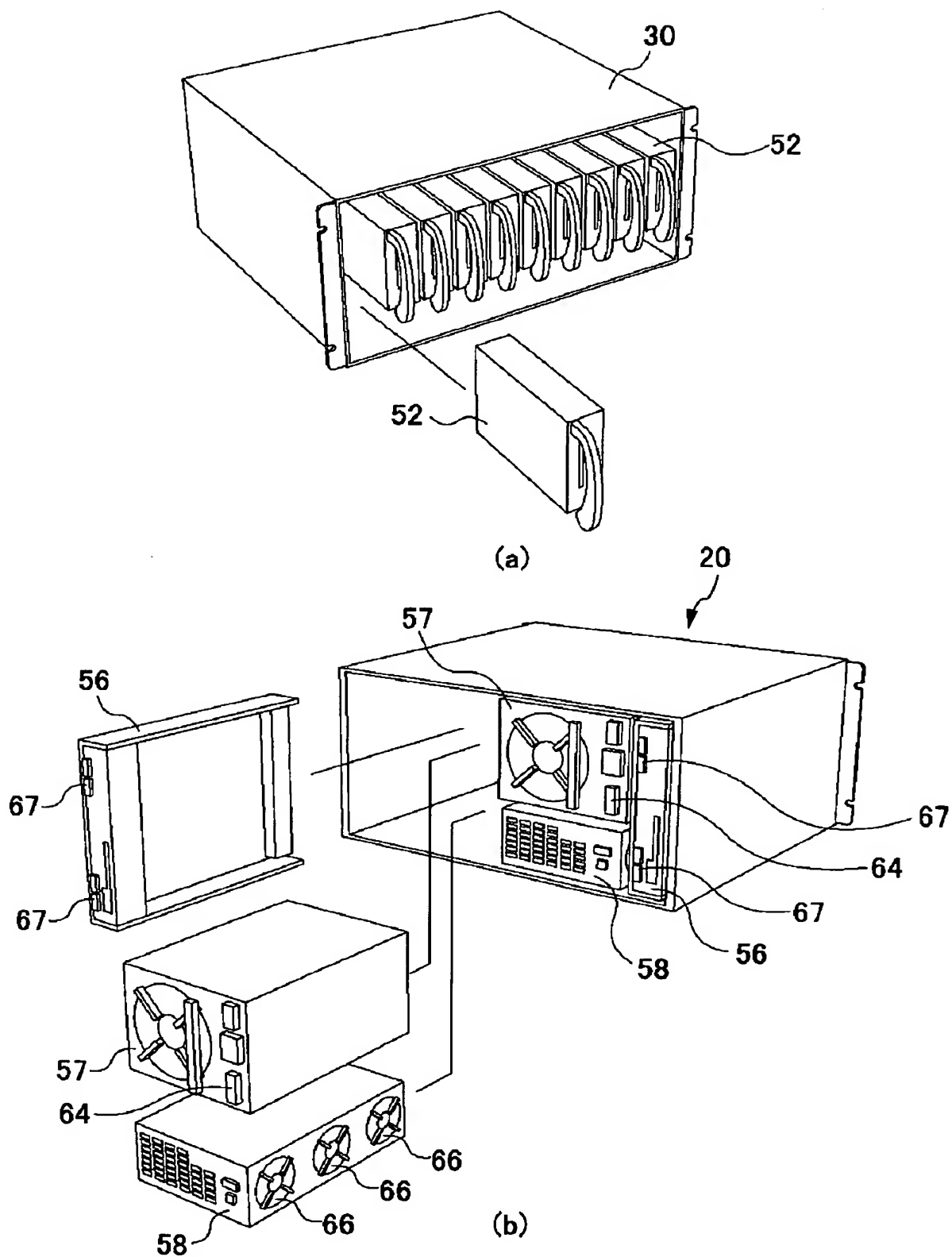
【図 1】



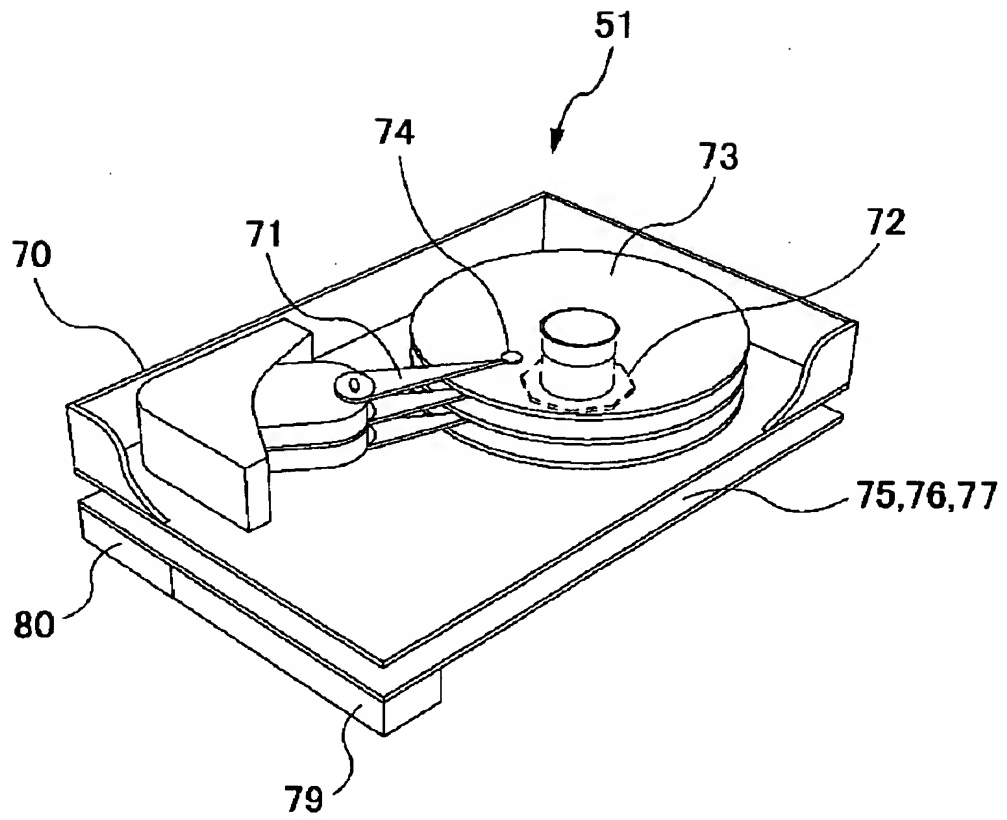
【図2】



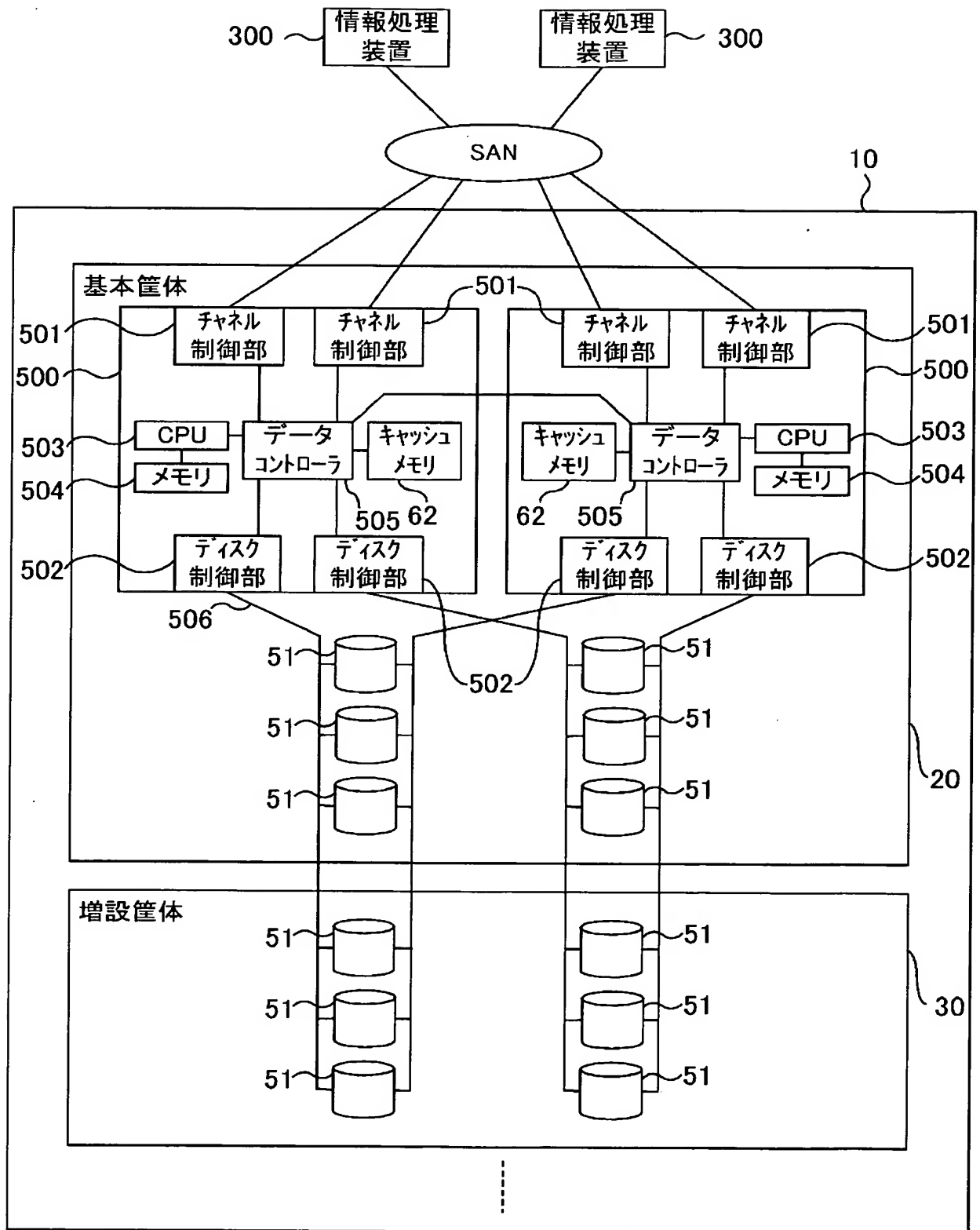
【図3】



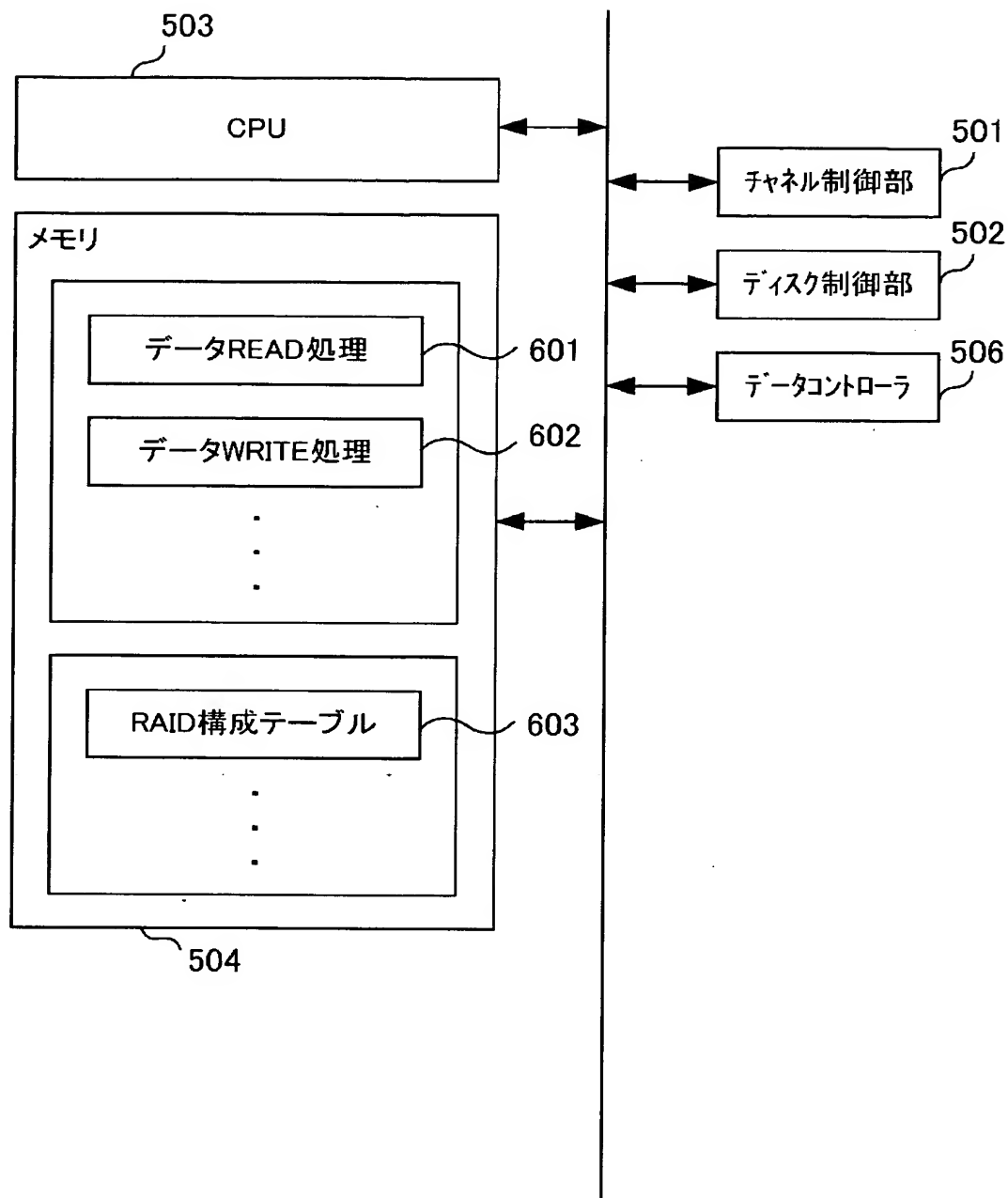
【図4】



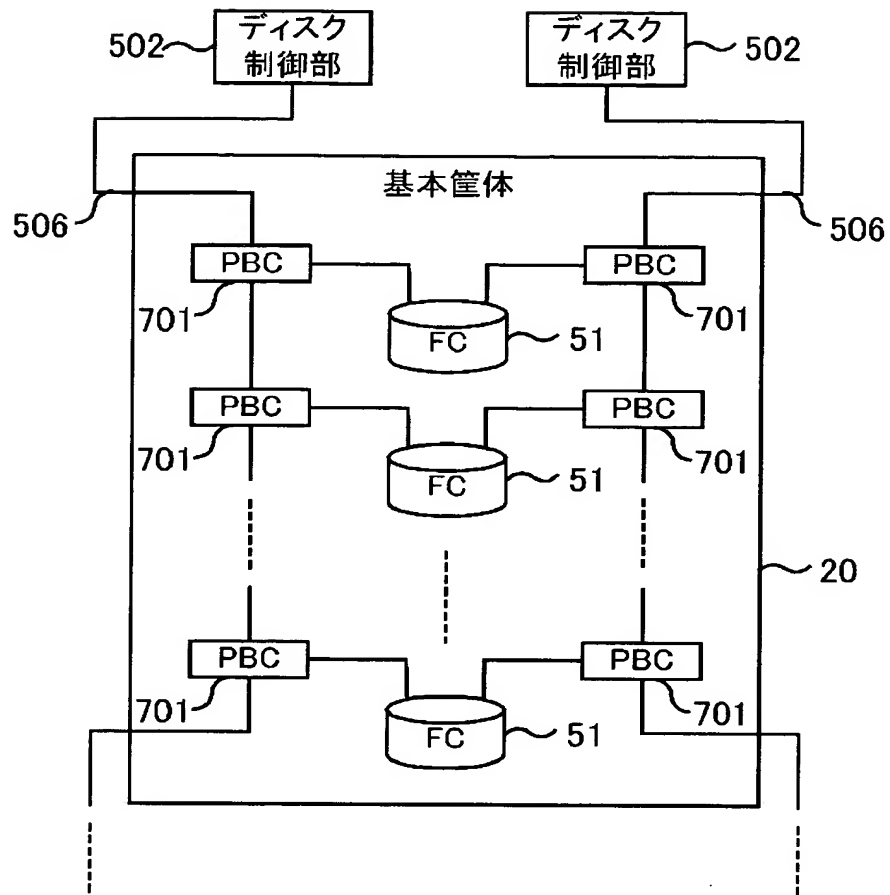
【図5】



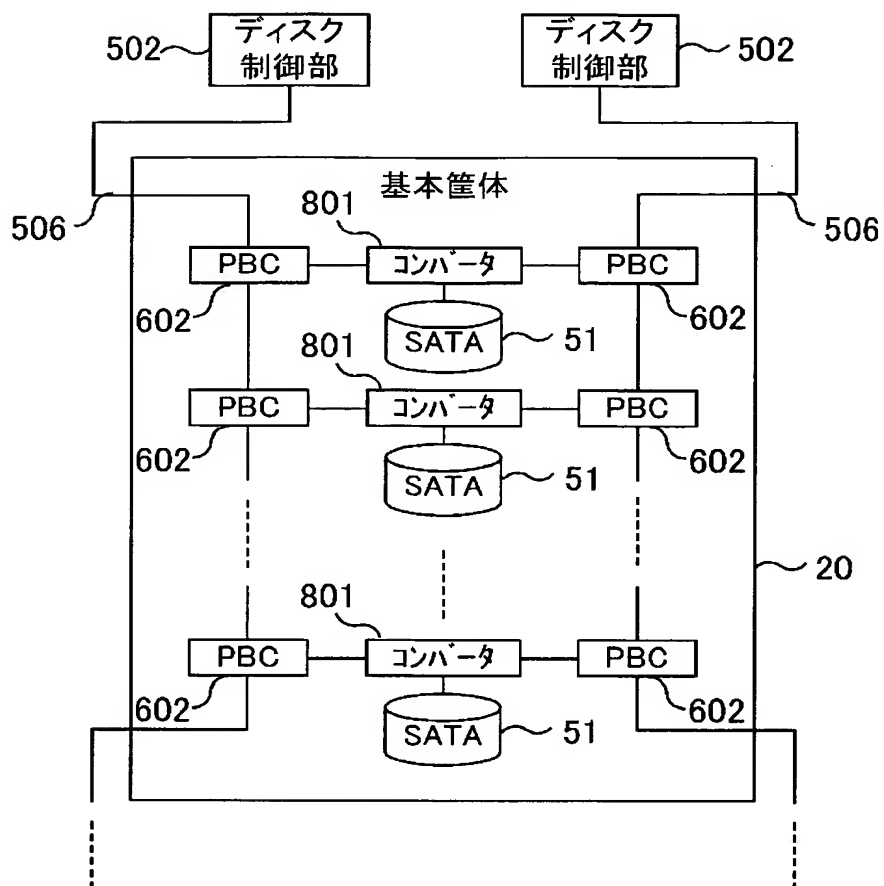
【図6】



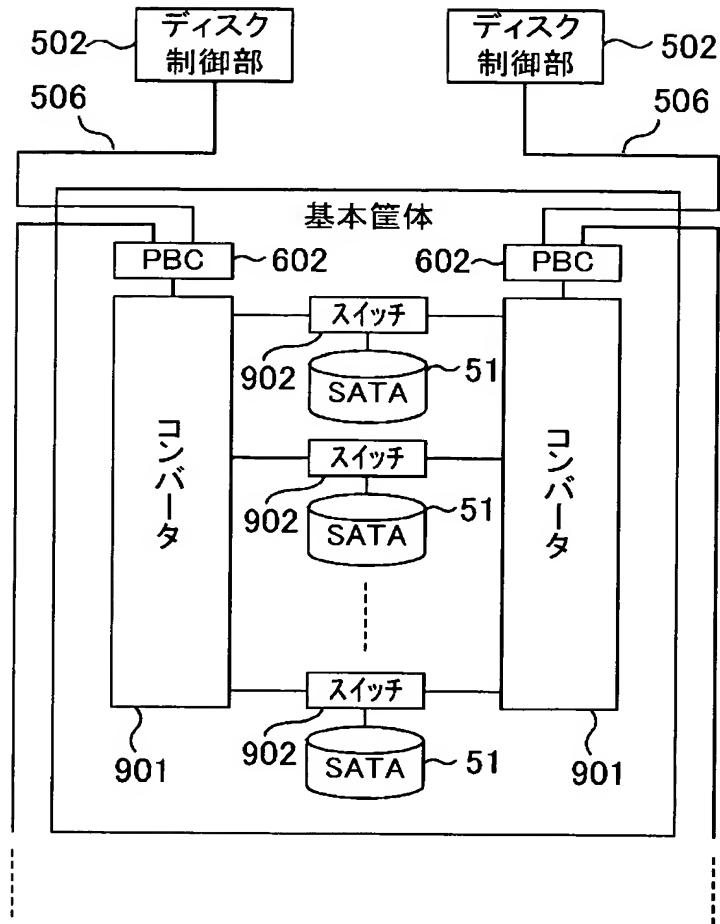
【図7】



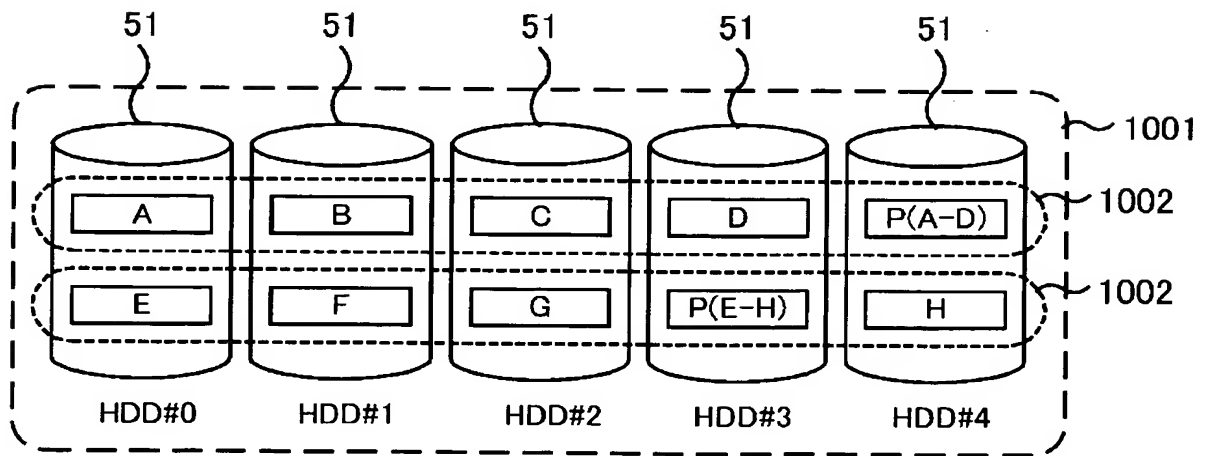
【図 8】



【図9】



【図10】

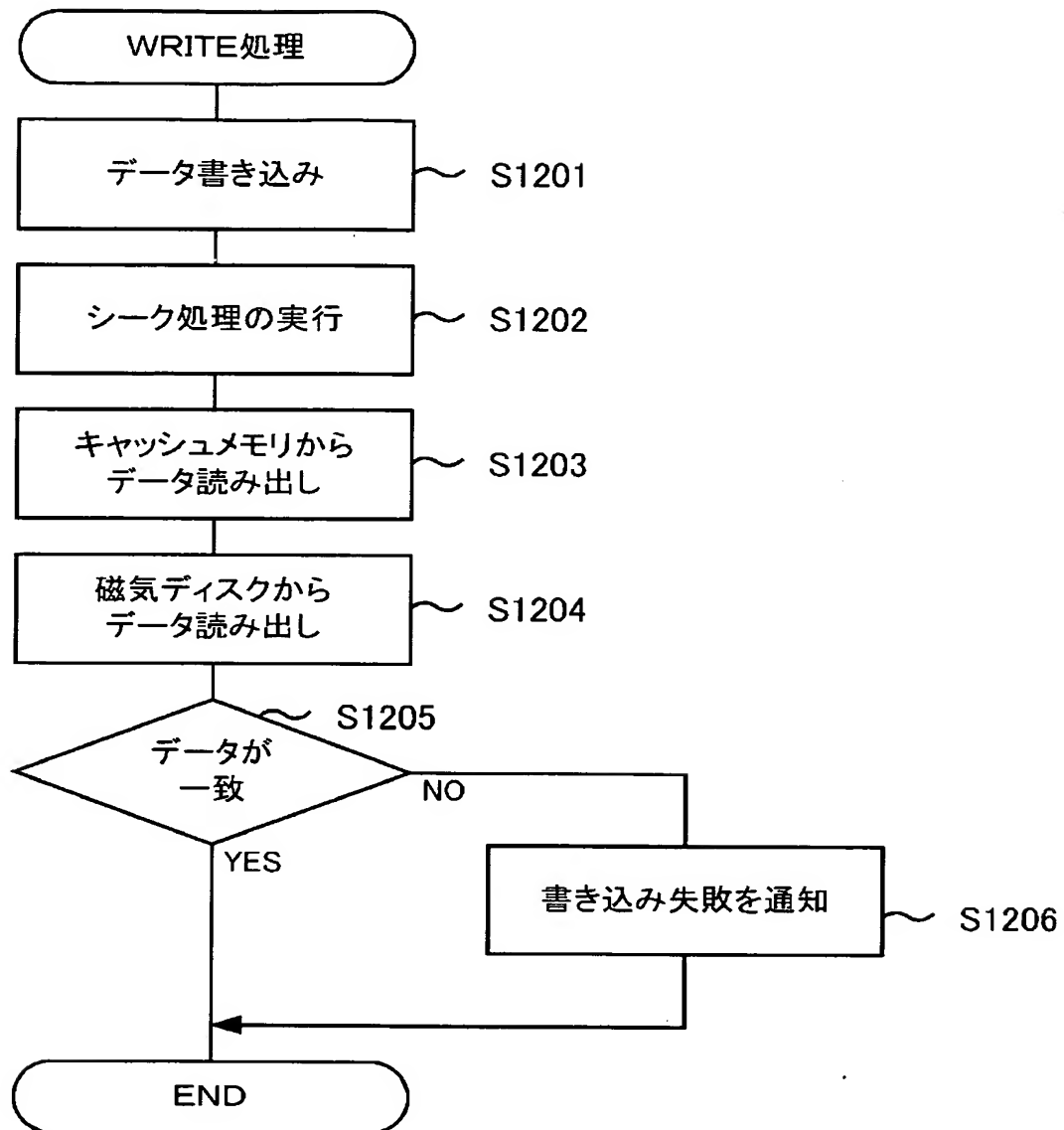


【図 11】

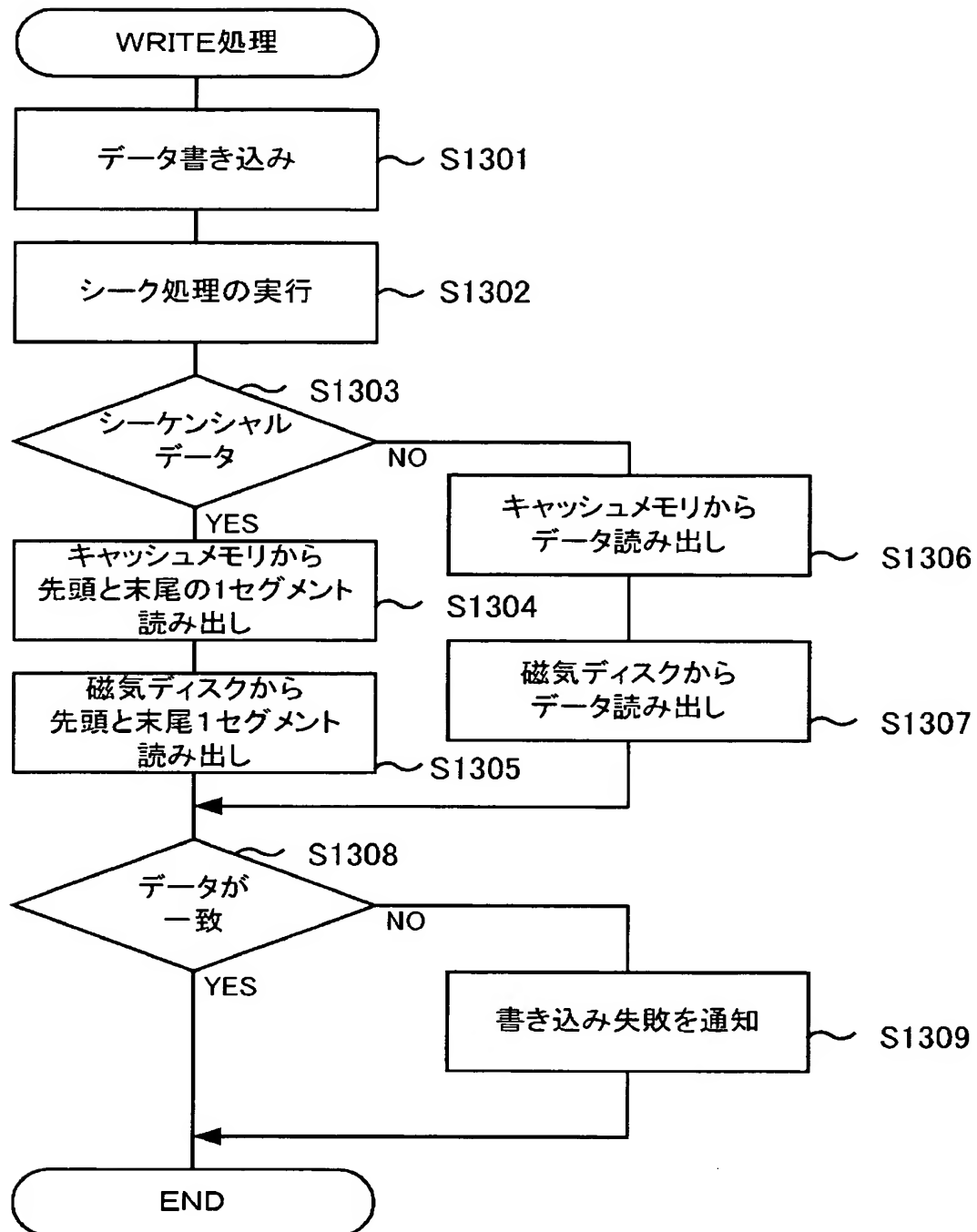
セクタ番号 ドライブ番号	LBA #1-128	LBA #129-256	LBA #257-384	...
HDD#0	0	0	1	...
HDD#1	1	0	0	...
HDD#2	0	1	0	...
...

1101

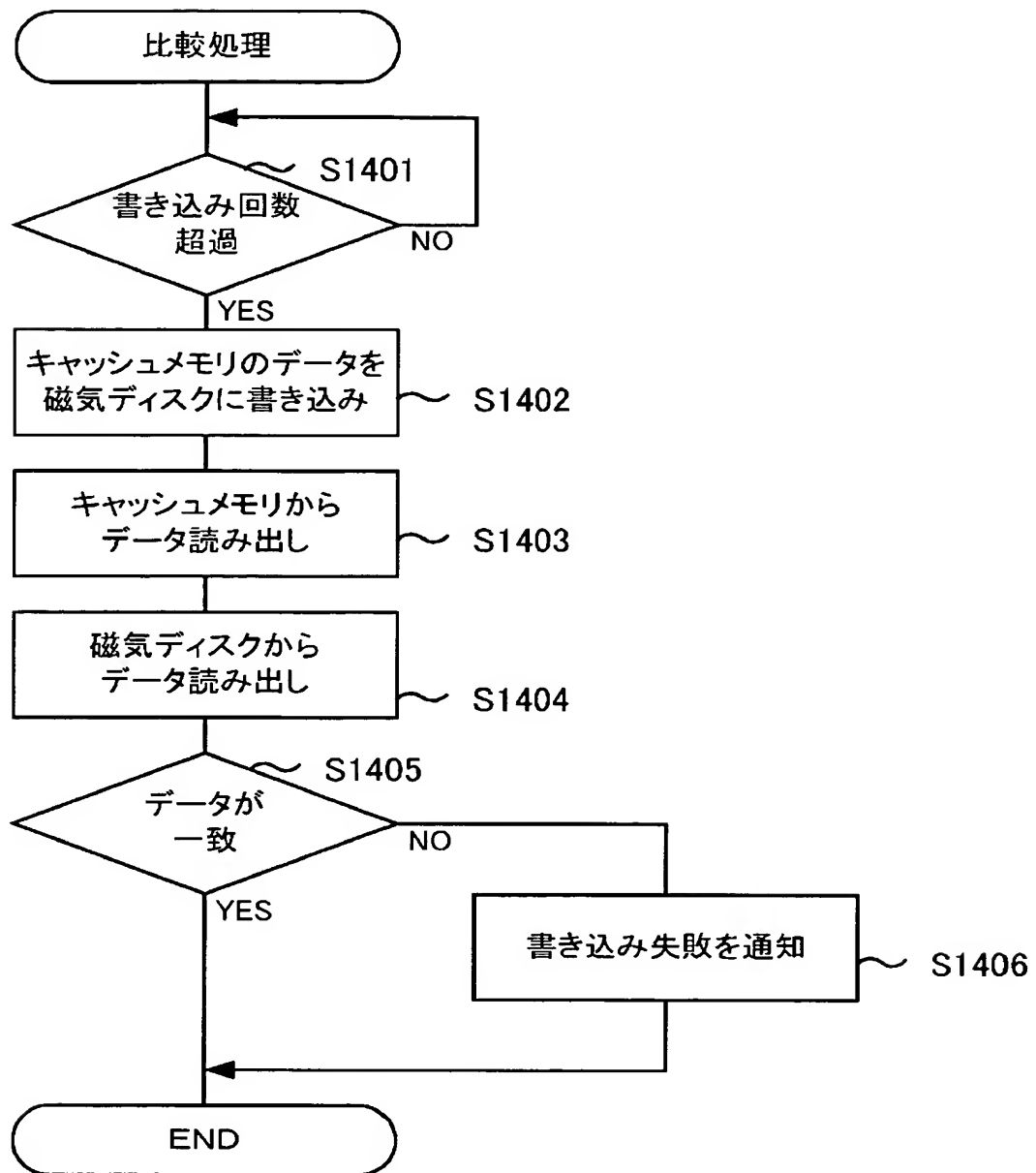
【図 12】



【図13】



【図14】

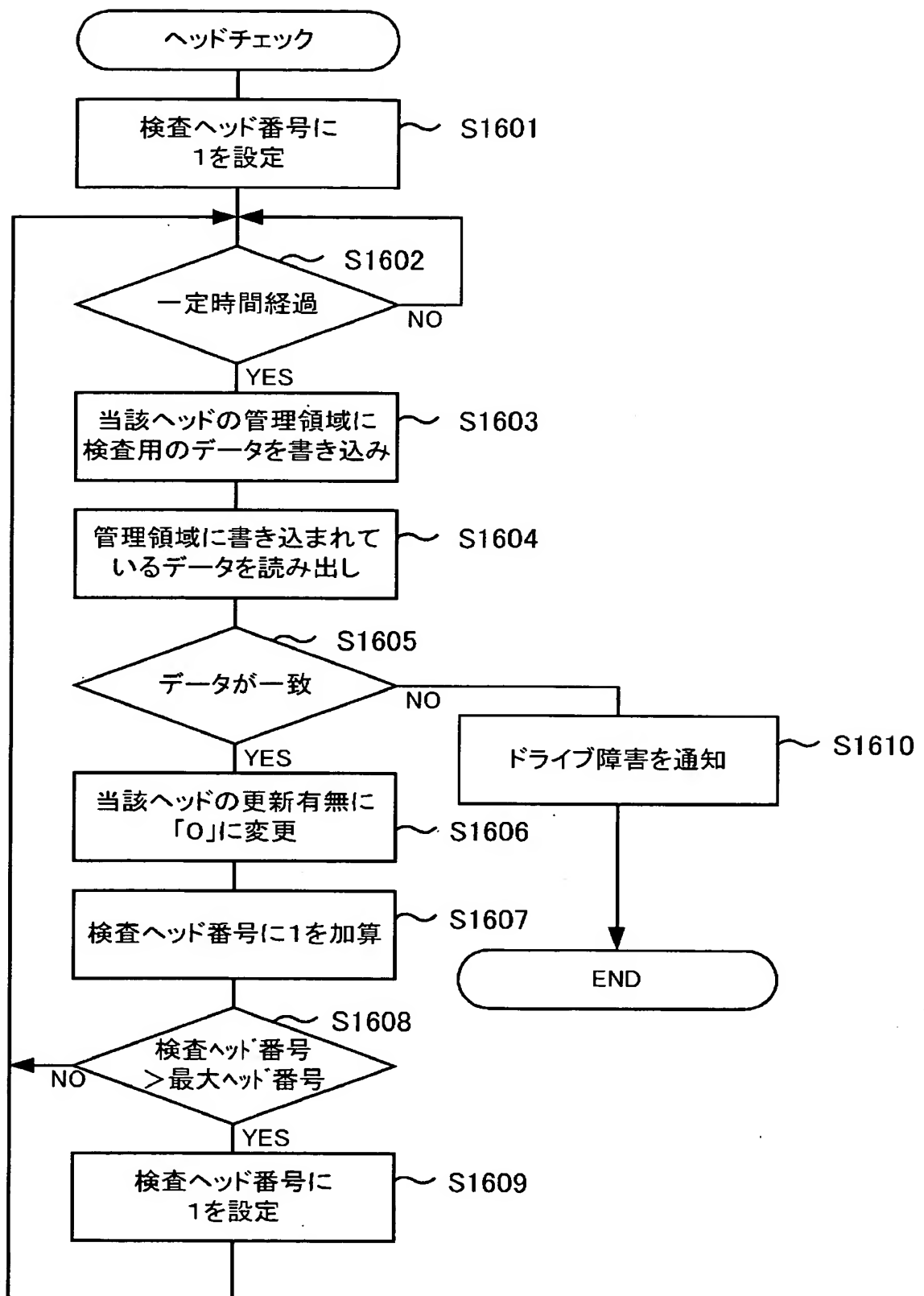


【図 1 5】

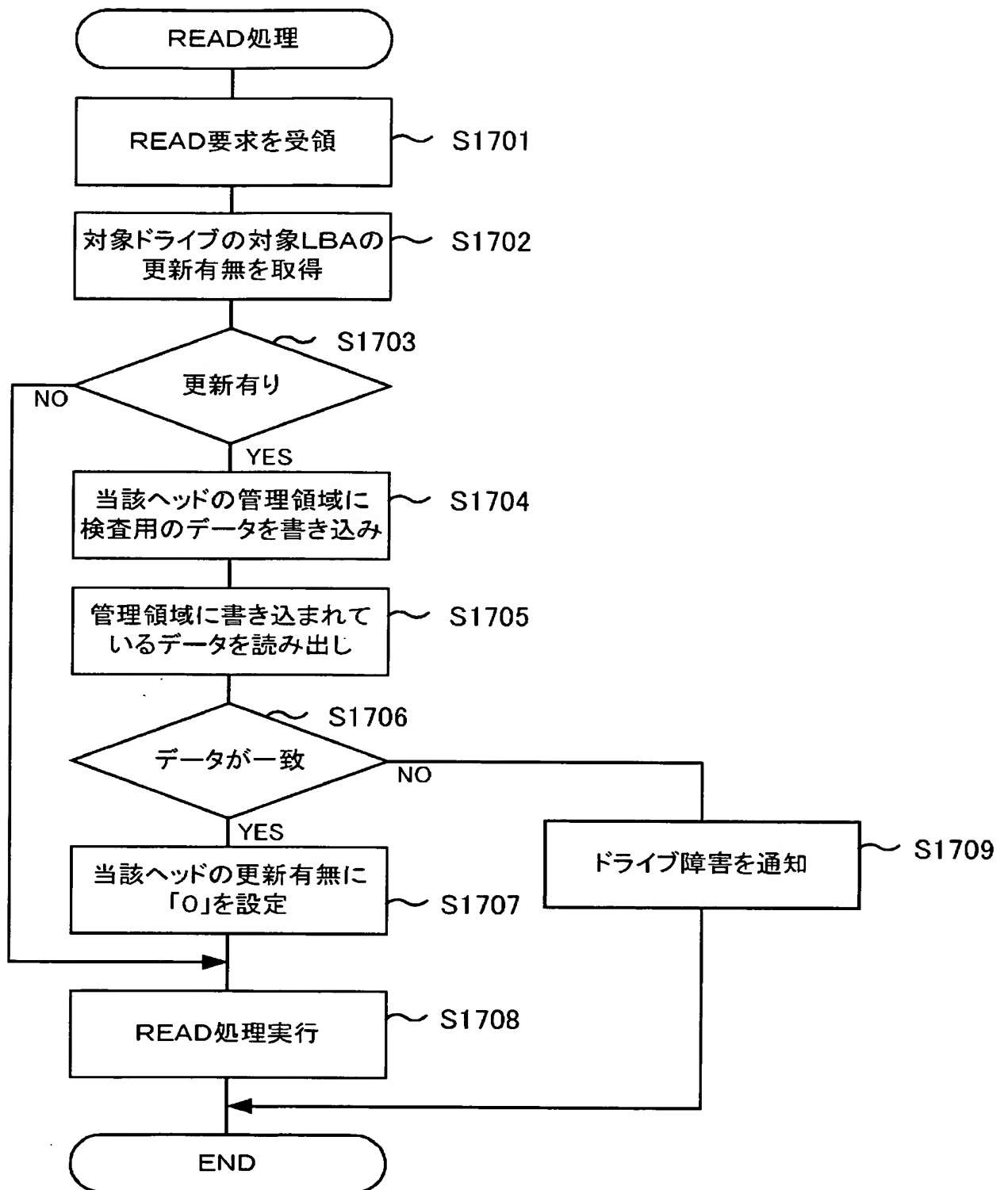
		LBA #1-128	LBA #129-256	LBA #257-384	...
HDD#0	ヘッド番号	#0	#0	#1	...
	更新有無	1	0	0	...
HDD#1	ヘッド番号	#0	#0	#1	...
	更新有無	0	1	0	...
...

1501

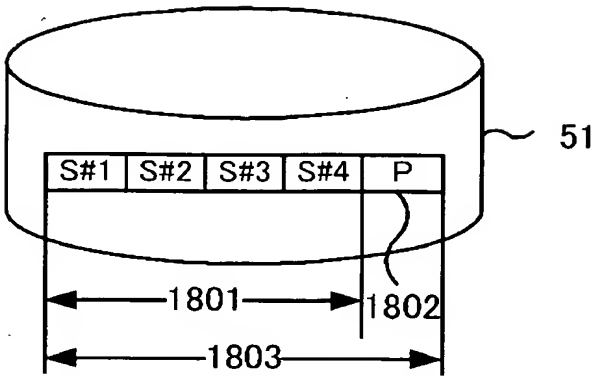
【図16】



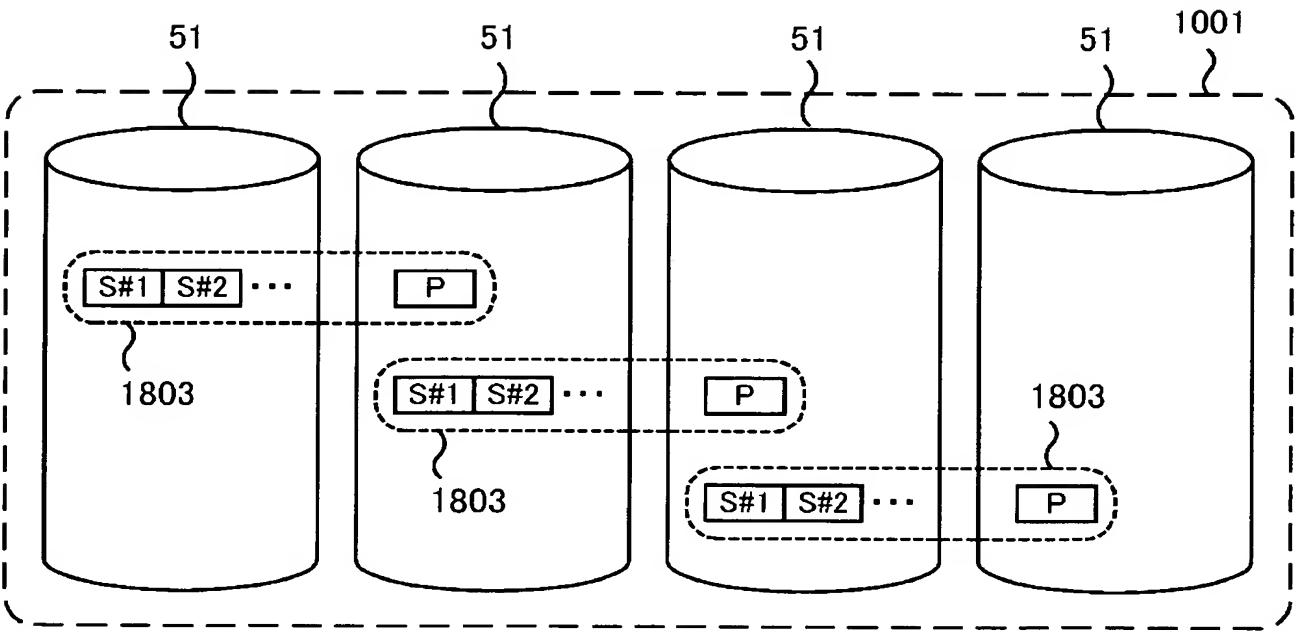
【図17】



【図 18】



【図 19】

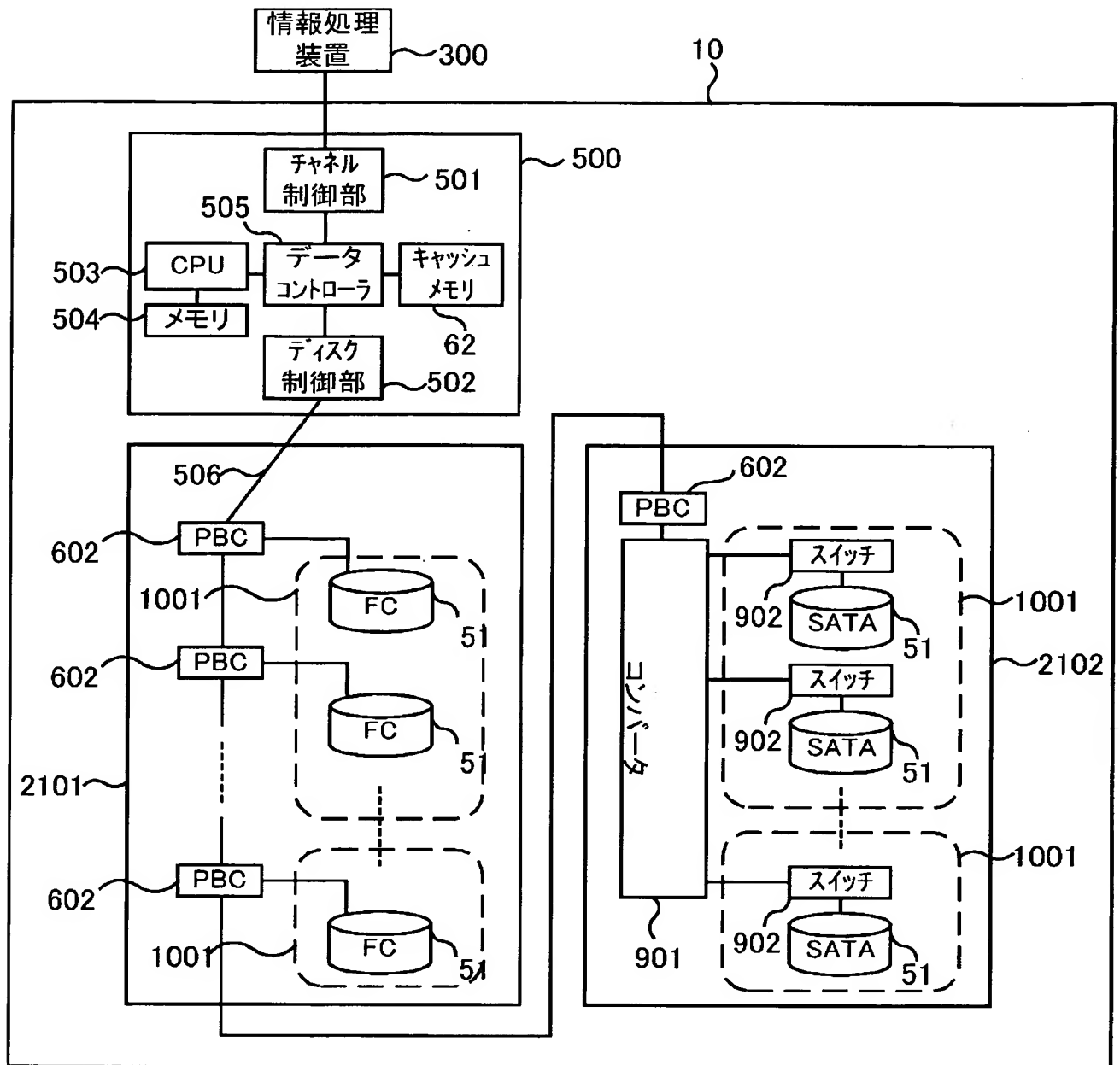


【図 2 0】

データ ユニット	ドライブ	ドライブLBA
000-129	#0	000-064
	#1	000-064
130-259	#2	000-064
	#0	065-129
⋮	⋮	⋮

2001

【図 21】



3 【書類名】 要約書

【要約】

【解決手段】 ファイバチャネルのハードディスクドライブが収容されている第一の筐体と、シリアル A T A のハードディスクドライブが収容されている第二の筐体と、第一の筐体と第二の筐体とを制御するコントローラとを有し、コントローラは、第二の筐体のシリアル A T A のハードディスクドライブに記憶されているデータを読み出す際に、当該データが記憶されている前記ハードディスクドライブが属している R A I D グループの全ての前記ハードディスクドライブから、当該データを含む複数のデータと当該複数のデータに対するパリティデータとを読み出し、当該データを含む複数のデータが誤った内容でハードディスクドライブに書き込まれていないか検査する。

【選択図】 図 2 1

特願 2 0 0 3 - 4 0 0 5 1 7

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所